MECHANISMS OF BINDING DIVERSITY IN PROTEIN DISORDER: MOLECULAR RECOGNITION FEATURES MEDIATING PROTEIN INTERACTION NETWORKS

Wei-Lun Hsu

Submitted to the faculty of the University Graduate School
in partial fulfillment of the requirements
for the degree
Doctor of Philosophy
in the Department of Biochemistry and Molecular Biology,
Indiana University

July 2013

Accepted by the Faculty of Indiana University, in partial
fulfillment of the requirements for the degree of Doctor of Philosophy.

A. Keith Dunker, Ph.D., Chair

Yaoqi Zhou, Ph.D.

Doctoral Committee

Thomas D. Hurley, Ph.D.

April 23, 2013

Vladimir N. Uversky, Ph.D.

ACKNOWLEDGEMENTS

I would like to take the opportunity to thank all the people who provided me with their help and support. I fully appreciated what they have done for me.

I would like to give my sincere gratitude to my adviser, Dr. A. Keith Dunker for his unreserved support and patient instruction during the past few years. His passion in research and outstanding accomplishment in science inspire me in many aspects. The great enthusiasm to the academic society he has especially makes me ways. Under Keith's guidance, I learned and was trained to combine bioinformatics analysis and laboratory experimentation to do intrinsically disordered protein research, which gives me a broad view to evaluate complicated biological questions in a systematic way. I really appreciate all the help Keith offered while I was in the most difficult time in my life. Without his support, I could not accomplish my dream to study in the U.S. In the meanwhile, Keith is also a good instructor to train and encourage students to develop their own innovative ideas and figure out solutions independently. He helped a lot to shape me and show me how to approach problems. I am so lucky to have Keith as my mentor that I could have the chance to explore my research interests, broaden my skill set and figure out my future career plan upon completion of my Ph.D. study.

I also want to thank my research committee, Dr. Vladimir N. Uversky, Dr. Yaoqi Zhou, Dr. Thomas D. Hurley and Dr. Pedro Romero for their valuable suggestions and comments to help develop my thesis work. I would also like to show my thankfulness to the Biochemistry and Molecular Biology department for continuing supporting in students' research and career development. I appreciated all the assistance from other

faculty members in our department as well, including Dr. Georgiadis, Dr. DePaoli-Roach, Dr. Goebl, Dr. Meroueh, Dr. Zhang, Dr. Wek, Dr. Hoang and Dr. Takagi.

In addition, I want to say thanks to all the members in Dr. Dunker's laboratory. Without their support, I can't accomplish what I have done. Thank you, Chris, Jingwei, Bin, Eshel, Caron, Fei, Maya and Bo for always being my technical and mental support. I also appreciated the chance to collaborate with other researchers outside of Indiana University. I thank Dr. Sarah Bondos and Hao-Ching Hsiao at Texas A&M University for sharing their fantastic work regarding to partner selection of Ubx protein, Dr. Lukasz Kurgan and Fatemeh Miri Disfani at the University of Alberta for their development of the MoRFpred disordered binding site predictor, Dr. Gil Alterovitz and Jonah Kallenbach in Harvard Medical School for working together to construct the MoRF-partner binary predictor.

Finally, I want to thank Yayue, Yunlong, Fucheng, Baohua, Hongying, Wenyan, Sue, Shelly, Yan, Yanlu, my family and friends for their endless support. Thank you all!

PREFACE

To innocence, and curiosity…

ABSTRACT

Wei-Lun Hsu

Mechanisms of Binding Diversity in Protein Disorder: Molecular Recognition Features Mediating Protein Interaction Networks

Intrinsically disordered proteins are proteins characterized by lack of stable tertiary structures under physiological conditions. Evidence shows that disordered proteins are not only highly involved in protein interactions, but also have the capability to associate with more than one partner. Short disordered protein fragments, called "molecular recognition features" (MoRFs), were hypothesized to facilitate the binding diversity of highly-connected proteins termed "hubs". MoRFs often couple folding with binding while forming interaction complexes. Two protein disorder mechanisms were proposed to facilitate multiple partner binding and enable hub proteins to bind to multiple partners: 1. One region of disorder could bind to many different partners (one-to-many binding), so the hub protein itself uses disorder for multiple partner binding; and 2. Many different regions of disorder could bind to a single partner (many-to-one binding), so the hub protein is structured but binds to many disordered partners via interaction with disorder. Thousands of MoRF-partner protein complexes were collected from Protein Data Bank in this study, including 321 one-to-many binding examples and 514 many-to-one binding examples. The conformational flexibility of MoRFs was observed at atomic resolution to help the MoRFs to adapt themselves to various binding surfaces of partners or to enable different MoRFs with non-identical sequences to associate with one specific

binding pocket. Strikingly, in one-to-many binding, post-translational modification, alternative splicing and partner topology were revealed to play key roles for partner selection of these fuzzy complexes. On the other hand, three distinct binding profiles were identified in the collected many-to-one dataset: similar, intersecting and independent. For the similar binding profile, the distinct MoRFs interact with almost identical binding sites on the same partner. The MoRFs can also interact with a partially the same but partially different binding site, giving the intersecting binding profile. Finally, the MoRFs can interact with completely different binding sites, thus giving the independent binding profile. In conclusion, we suggest that protein disorder with post-translational modifications and alternative splicing are all working together to rewire the protein interaction networks.

A. Keith Dunker, Ph.D., Committee Chair

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

MoRF	Molecular Recognition Feature
IDP	Intrinsically Disordered Protein
NMR	Nuclear magnetic resonance
ANS	1-Anilino-8-naphthalene-sulfonate
PTM	Post Translational Modification
IDR	Intrinsically Disordered Region
ASE	Alternative Splicing Event
ELM	Eukaryotic Linear Motif
LM	Linear Motif
SLiM	Short Linear Motif
RISP	Regions of Increased Structural Propensity
PDB	Protein Data Bank
RMSD	Root Mean Square Deviation
OR	Overlap Ratio
CI	Confidence Interval
SCOP	Structural Classification of Proteins
NR	Nuclear Receptor
PPI	Protein-Protein Interaction
UniProt	Universal Protein Resource
iMoRF	Immune-Related MoRF

CHAPTER 1

Introduction

1.1. Intrinsic Protein Disorder and Protein Functions

Intrinsically disordered proteins (IDPs) are a group of proteins that lack stable tertiary structures either partially or in their entirety. Their structural conformations are too dynamic to be described by a single conformation under physiological conditions. IDPs still can be identified by more than 40 experimental methods, such as x-ray crystallography (missing density), Nuclear magnetic resonance (NMR) (lack of chemical dispersion in 1H-15N NOEs), far-UV (170-250nm) circular dichroism (lack of secondary structure), protease sensitivity (readily cleaved by proteases), 1-Anilino-8-naphthalene-sulfonate (ANS) binding (lack of hydrophobic cores) and so on. Protein disorder has been found to exist in nature as disordered tails, linkers, domains, or entirely unfolded as collapsed or extended forms (Figure 1) [1]. The existence of IDPs challenge the traditional biochemistry view of sequence-structure-function paradigm since these proteins still carry out important biological functions without well-defined structures. In other words, the structure of a protein may not always define its function or a single unique structure cannot describe their function. However, in some cases, these disordered regions can adopt specific three dimensional structures after binding to another molecule. There are some possible reasons why IDPs lack stable structures. Some researchers believe IDPs are unstructured only when lacking a ligand/partner or other factors that promote their folding, but others, including our laboratory, believe IDPs' lack of structure is encoded by their amino acid sequences just like structured proteins.

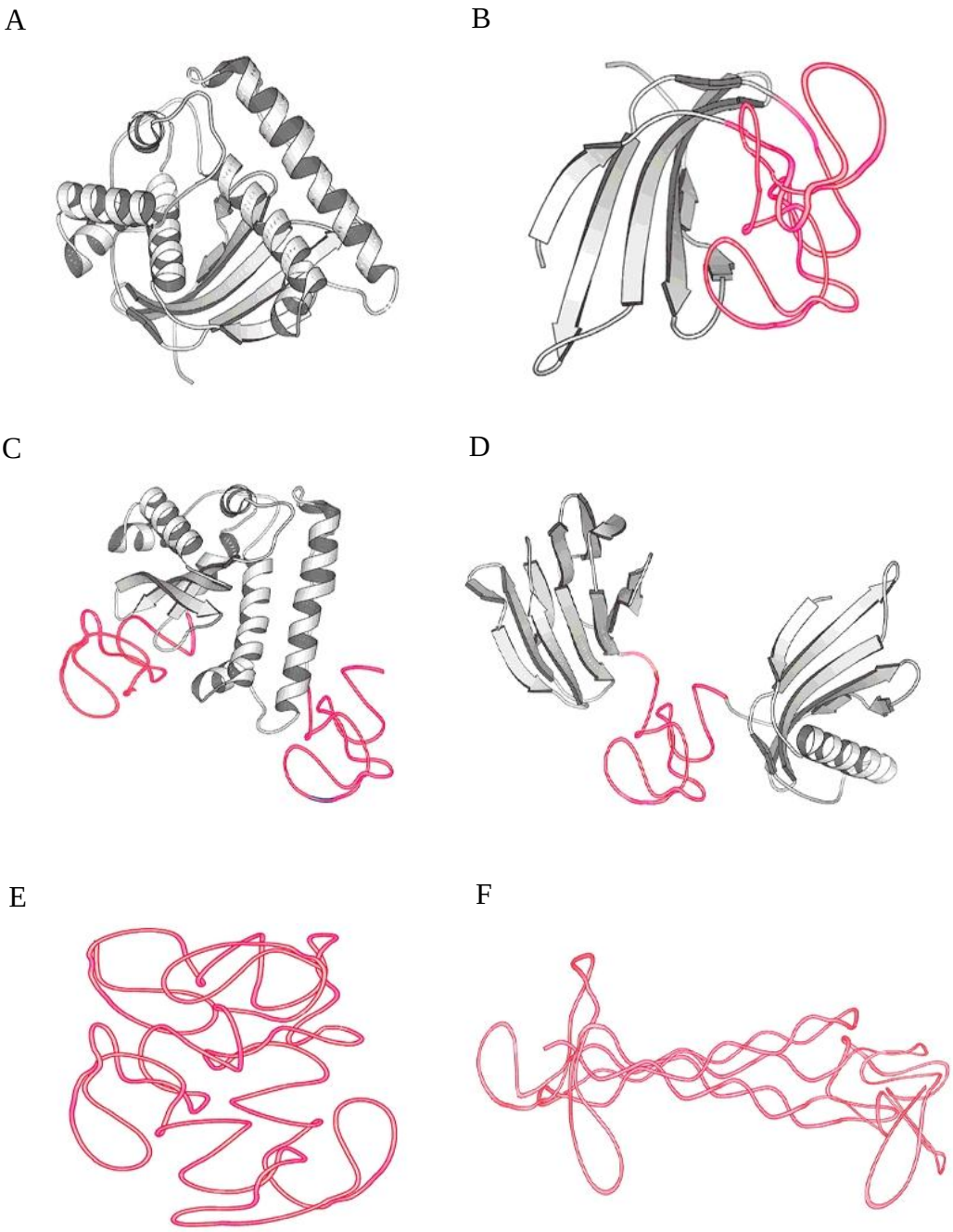


Figure 1. Various forms of protein structures: (A) structured domain, (B) disordered domain, (C) disordered tails, (D) disordered linker, (E) collapsed disorder and (F) extended disorder. Red parts of structures imply disordered regions. The diagram is adapted from DisProt Database [1].

IDPs are often referred to using alternative names, such as naturally unfolded proteins, intrinsically unstructured proteins, flexible/dynamic proteins, conformational disorder, extended polypeptide, mobile domains, molten globule, random coils or disordered proteins. Genomics and proteomics studies have revealed protein disorder is highly abundant in various organisms, such us in humans and viruses. Eukaryotes generally have higher intrinsically disordered contents than prokaryotes. A quantitative and qualitative measurement of the extent of protein disorder in 3484 species with known genomes was performed by Xue et al. [2]. Viruses were found to have the widest spread of disorder content (from 7.3% in human coronavirus NL63 to 77.3% in avian carcinoma virus) in their study.

Several studies have revealed the possibility of the hypothesis: protein disorder is used for signaling because of its unique structural properties. Many bioinformatics studies claim that disordered proteins involve more in signaling pathway, gene regulation, molecular recognition and cell control particularly while structured proteins often involve in catalysis, membrane transport and small molecules binding [3-7].

Many biological events in which disordered proteins participate are found to be regulated by post translational modifications (PTMs) and alternative splicing events (ASEs) [8,9]. Fukuchi et al. explored a variety of protein modification events in different subcellular localizations and found protein disorder are highly enriched in nuclear proteins (47%) compared to mitochondria proteins (13%) [8]. Also, phosphorylation and O-linked glycosylation sites were frequently observed to localize in intrinsically disordered regions (IDRs). They suspected the O-linked glycans are attached to IDRs in order to protect the protein from proteolytic cleavage in the extracellular environment.

Besides PTMs, alternative splicing events (ASEs) have been associated with IDRs by various laboratories [8,9].

1.2. Intrinsic Protein Disorder in Protein-Protein Interactions

Many proteins execute their biological functions through protein-protein interactions. By binding to interacting partners, proteins can deliver signals to other molecules. For example, hormone neurotransmitters and their receptors trigger various signal transduction pathways following their mutual interaction, antibody recognition of peptide antigens leads to B-cell activation, and the interaction between G-protein coupled receptors and G-proteins leads to the transduction of many biological signals.

Protein-protein interaction networks underlie a wide variety of biological functions, ranging from regulating cell division to responding to external signals. High throughput methods have enabled researchers to map out sets of protein-protein interactions over entire proteomes. Mapping protein-protein interactions leads to networks that are far from random. While most proteins have only a few interacting partners, the studies reveal complex networks in which a small number of proteins, called hubs, are observed, to have multiple interacting partners. Indeed, in some cases hubs bind to 15, 20, 50 or even more partner proteins. As expected for such network architecture, deletion of a protein with only a few partners is typically less deleterious than the deletion of a hub protein [10,11].

How do such networks arise from simpler precursors? Other networks of a similar architecture arise because “the rich get richer”; units with more connections have a higher probability of adding even more connections over time as compared to the units with fewer connections. This suggests that highly connected proteins have special

features that facilitate their binding to multiple partners and that facilitate binding to new partners that arise through mutation [12]. What are these special features?

Theoretical arguments [13,14] and experimental data [15,16] suggest that unfolded or disordered protein can very readily change shape and thereby easily adapt to multiple, distinct partners. The common involvement of disorder in hub proteins' interactions has been supported by several subsequent studies [17-19]. Intrinsically disordered proteins often bind to more than one partner. Thus, we proposed that the special feature of hub proteins enabling their binding to multiple partners is likely to be intrinsic disorder. In support of IDPs as being important for binding to multiple partners, both hub proteins and their binding partners are observed to be enriched in disorder [19-21], and many additional studies support these concepts [17,22-31].

1.3. Characterization of Molecular Recognition Features (MoRFs) and their Binding Partners

With regard to IDP regions involved in binding, various descriptors have been used, such as eukaryotic linear motif (ELMs) [32,33], linear motifs (LMs) [34], short linear motif (SLiMs) [35,36], regions of increased structural propensity (RISPs) [37], and molecular recognition features (MoRFs) [38]. All of these describe similar phenomena, despite different approaches used by the various researchers for identification of binding segments. The identification of ELMs, LMs, or SLiMs start from sequence pattern or motif-based approaches, whereas the identification of RISPs and MoRFs start from short regions with binding indicators located within longer regions of predicted disorder. The motif-based and algorithmic approaches show significant overlap in their identification of their binding sites [34], suggesting that the different approaches associated with the

different names are merely emphasizing different aspects of the same types of binding interactions.

Because ELMs, LMs, and SLiMs all involve sequence motifs, these binding regions can be identified by simple pattern recognition methods, albeit with a high error rate due to their typically short length involving just a few key residues. Predicting protein-protein interaction sites in proteins can be used to supplement experimental approaches [39,40]. Predicting binding sites by sequence matches to the motifs of ELMs [32,33], LMs [34], SLiMs [35,36], or other collections of sequence patterns [41-43] provides one strategy for identifying potential binding sites located within IDPs or IDP regions. Using sequence characteristics that indicate short binding regions within longer regions of disorder offers a second strategy that does not depend on specific motifs, and several predictors have been developed that use this second strategy [44-48]. Such predictors have been used by experimentalists to help with the identification of binding regions within longer regions of disorder [37,49].

1.4. MoRFs in PDB: Their Length, delta ASA and Secondary Structures

Table 1 lists the number of MoRFs we collected in each filtering step in our 2008 and 2012 datasets. The criteria we used for screening MoRFs are slightly different in two aspects: the length of MoRF partners and the exact sequence we use for sequence alignment. Basically, the MoRF dataset grew about 2.7 folds over the past 4 years.

Table 1. Description of MoRF datasets built in 2008 and 2012.

Data set	March 2008	June 2012
Initial MoRF dataset (5-25)	4289	8084
MoRF dataset with biological interaction (>400Å^2)	3837	7064
MoRF dataset with globular partner (>70 vs. > 40)	3148	6171
MoRFs mapped to UniProt (ATOM vs. SEQRES)	1805	4839

The following Figures (2-4) give us a general overview of our 2008 MoRF dataset (4289 complexes) on MoRF length, surface area change upon binding (ΔASA) and secondary structure.

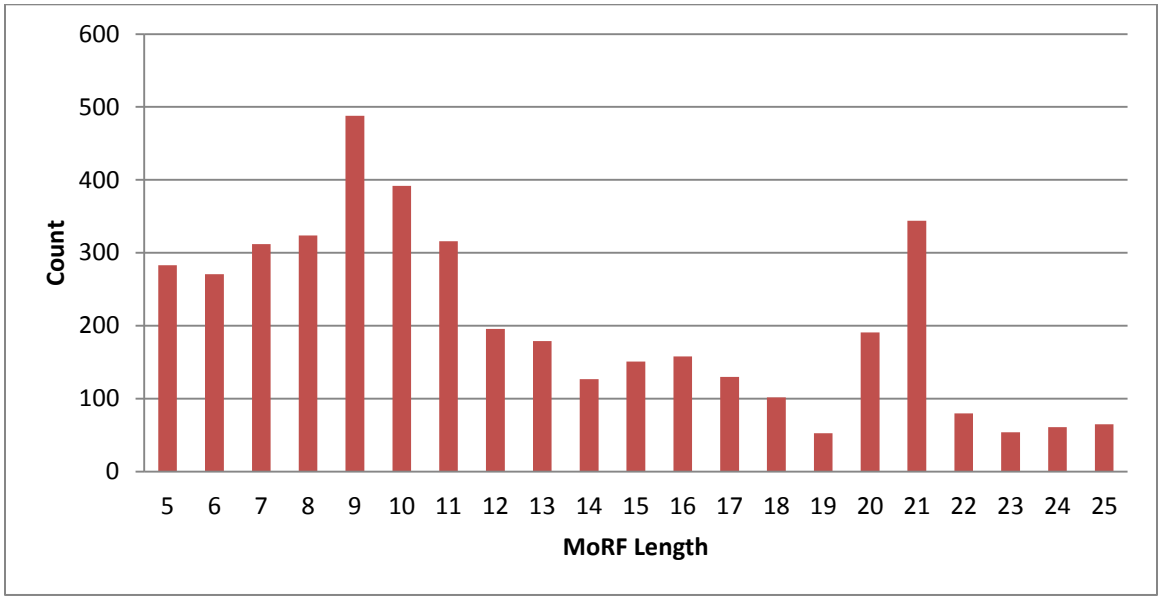


Figure 2. A histogram of MoRF length of the 2008 MoRF dataset.

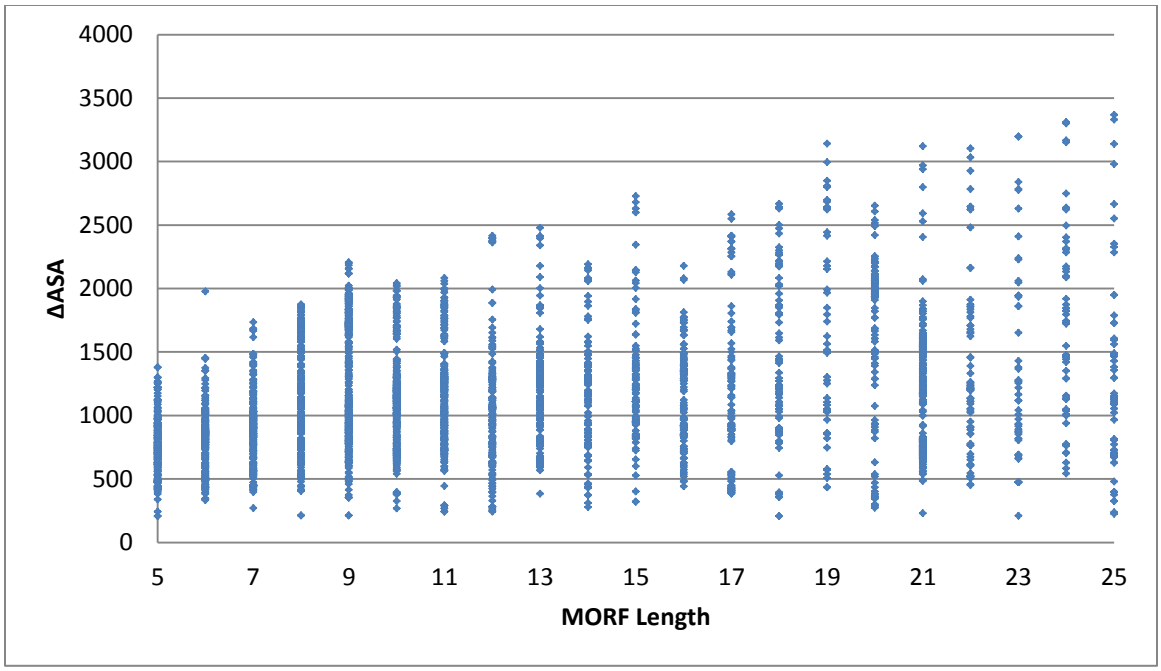


Figure 3. A scatter plot reveals a positive but not significant correlation between MoRF length and surface area change (ΔASA) upon binding.

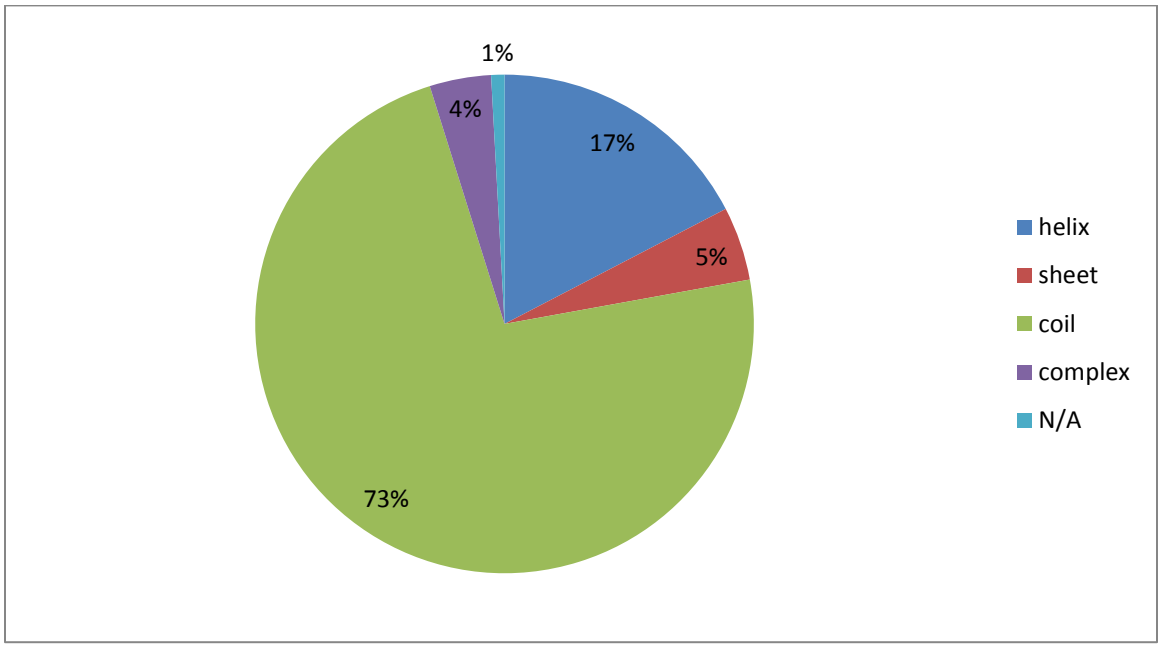


Figure 4. A pie chart of different MoRF types based on their secondary structures.

1.5. Validation on MoRFs (Gunasekaran-Tsai-Nussinov Graph)

Gunasekaran et al. developed a protocol [50] that we modified [38] to indicate whether a MoRF is likely to be disordered when unbound. The Gunasekaran-Tsai-Nussinov graph provides a scale that measures confidence with which one can say whether a protein is ordered or disordered. The farther the point, which corresponds to a given chain, is from the dividing black line (boundary), the greater the confidence with which a protein can be classified into either of the classes. Points above the line correspond to disordered chains like Figure 5 shows below. All the 842 MoRFs selected form our 2008 MoRF dataset (a non-redundant set) are validated as likely to be disordered before the binding events.

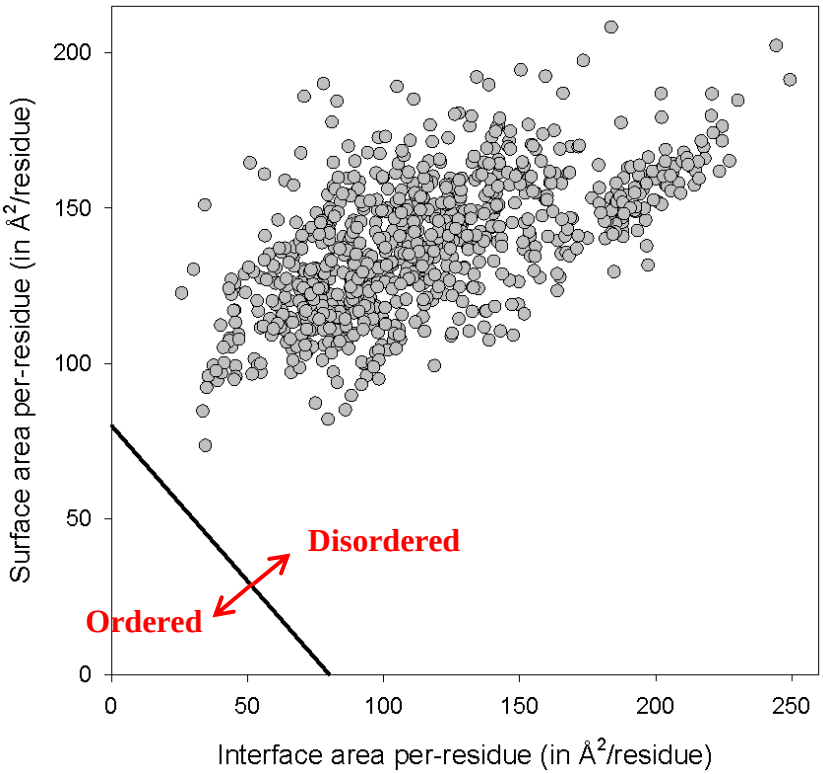


Figure 5. A Gunasekaran-Tsai-Nussinov graph example (adapted from Bioinformatics **28**, i75-83).

1.6. Two MoRF Mechanisms in Hub Proteins

We further suggested two ways that disorder could be used by hub proteins for binding to multiple partners: 1. One region of disorder could bind to many different partners (one-to-many binding), so the hub protein itself uses disorder for multiple partner binding; and 2. Many different regions of disorder could bind to a single partner (many-to-one binding), so the hub protein is structured but binds to many disordered partners via interaction with disorder [51]. Since this initial proposal, we [19,22,23] and many others [20,21,24-31,52] have provided additional evidence that hubs and/or their binding partners are especially enriched in intrinsic disorder, with both the many-to-one and one-to-many processes involving the use of intrinsic disorder.

The C-terminal region of p53 uses disorder to bind to more than 45 different proteins and to form a tetramer, but only six of these complexes and the tetramer have had their structures deposited in the Protein Data Bank (PDB) [46]. One particular p53 segment "SHLKSKKGQSTSRHKKLMFKTE" (residues 367-388), which is both an ELM and a MoRF and which is located at the C-terminus, morphs into an α-helix when binding with S100ββ, into a β-sheet with sirtuin, into an irregular structure with CREB binding protein (CBP) and into another irregular structure with cyclin A2 as a partner [46].

Very different biological processes are transduced via these four different interactions involving the same segment of p53: The CDK2/cyclin A2 complex regulates progression of S phase of the eukaryote cell cycle by recognizing diverse but structurally constrained target sequences (KXL/RXL motif) from various substrates, including p53 [53]; deacetylase enzymes like the Sir 2 protein, which is a homologue of Sirtuin, can

lead to down-regulation of p53-dependent transcription by binding to the acetylated p53 peptide on lysine 382 [54]; the recognition of acetylated lysine 382 in p53 by the conserved bromo-domain of transcriptional coactivator CBP is very specific, leading to the recruitment of p53 acetylation-dependent coactivator following DNA damage and to the activation of cyclin-dependent kinase inhibitor p21 [55]; dimeric S100 calcium binding protein B can sterically block the phosphorylation and acetylation sites of on p53 that are critical for the activation important transcription; finally, the peptide derived from the region of p53 was found to undergo a disorder-to-order conformational change while binding to Ca2+ loaded S100ββ [56]. Thus, this same intrinsically disordered segment plays roles in a diverse set of signaling pathways.

The highly conserved 14-3-3 protein family has been reported to associate with over 200 different but mostly phosphorylated proteins [57]. Phosphorylation plays a central role in cellular regulation, either by altering a protein's activity directly or by inducing specific protein-protein interactions. Protein phosphorylation events are often coupled with domain-binding motifs, highlighting a potential switch-like function of phosphorylation. In part, the ability of 14-3-3 to associate with many different proteins is the result of its specific phospho-serine/phospho-threonine binding activity. These phosphorylation sites are often surrounded by disorder-promoting residues. From this observation, a bioinformatics study suggested that over 90% of the 14-3-3 protein partners do not adopt a defined three-dimensional structure in total or in part [58]. This implies structural disorder in 14-3-3 partners is the key characteristic for promoting this binding diversity. But how the 14-3-3 partners have diverged with respect to their primary structure and yet still maintain binding to 14-3-3 as an unanswered question.

In the 14-3-3 many-to-one binding example, 3D structures have been determined for five different complexes having different disordered sequences, namely a peptide fragment from the tail of histone H3, serotonin N-acetyltransferase (AANAT), a phage display-derived peptide (R18), and peptides described as motifs 1 and 2 (m1 and m2). All five of these peptides associate within a common binding groove in 14-3-3 [46]. Within the superimposed structures of the five peptides, the central three binding residues show little divergence in backbone locations, but the backbones become more separated as one moves away from the central phosphorylated (or negatively charged) residue. This divergence is loosely correlated with the sequence similarity. The standard deviation of ΔASA for the peptide binding residues also show either end of the central cleft have the most binding diversity. Restricted backbone variability in bound 14-3-3 structures suggests that a large conformational change in 14-3-3 is not necessary for multiple specificities, but some small adjustments at the ends of binding helices may be unavoidable. The circular variances of the dihedral angles of residue side chains indicate side chain rearrangements also help accommodate different peptide sequences.

The multiple intrinsically disordered phosphorylated proteins bound by 14-3-3 regulate a wide range of cellular targets [59]. The diverse cellular processes involving these interactions with 14-3-3 include signal transduction, cell cycle control, apoptosis, transcriptional regulation, cytoskeleton rearrangements, cell adhesion, chromosome maintenance, protein localization, protein trafficking, protein degradation, exocytosis, endocytosis, development and stress response [60]. Therefore, molecular recognition by 14-3-3 proteins highlights the emerging importance of using system-based approaches to understand signal transduction event at the network biology level.

Many other protein-protein interactions are also mediated by the same many-to-one binding mechanism. Well known examples include MoRFs that interact with SH3, SH2, PDZ and WW domains [61-63]. However, the true extent and diversity of MoRF-mediated interactions is largely unknown.

We know of only two atomic resolution comparisons of more than one IDP binding to the same partner: two different peptides binding to TAZ1 domain [64] and five different peptides binding to 14-3-3 [46,65].

Our initial work [19,22,23,51] on disorder and protein-protein interactions focused on single binding sites that used regions of disorder. To be more complete, it is worth mentioning that, in addition to the one-to-many and many-to-one mechanisms used by single sites of disorder for multiple partner binding, hub proteins can also use multiple binding domain repeats likely connected by flexible (disordered) linkers [20], or hubs can use multiple binding sites one after another in long regions of disorder as we recently discussed [66]. Of course these additional, multi-site mechanisms can be multiplexed via one-to-many and many-to-one mechanisms, thus leading to extremely complicated protein-protein interaction networks.

1.7. Importance of MoRF Mechanisms in Hub Proteins

Independent of their roles in hub protein interactions, intrinsically disordered proteins (IDPs) lack of specific structures provide the basis for important biological functions [67,68] such as signal transduction, cell regulation, molecular recognition, and many other functions [3-7,64,69,70]. Many of these disorder-utilizing biological functions depend ultimately on disorder-based protein-protein interactions. Thus, understanding the structural basis of protein-protein interactions involving IDPs is

important for a wide variety of biological functions, not just as the mechanistic basis for hub protein function.

Both a hub protein's ability to bind multiple partners and the general importance of protein-protein interactions suggest that the use of flexibility for partner binding by IDPs and IDP regions is of considerable interest. However, despite the importance of understanding how one disordered region can bind to more than one partner, there have been very few structural comparisons at the atomic resolution level, either for one-to-many binding examples or for many-to-one binding examples. For the latter, we know of only two atomic resolution comparisons of more than one IDP binding to a single partner: namely, two different peptides binding to the TAZ1 domain [64], and five different peptides binding to 14-3-3ζ [46]. With regard to the former, we likewise know of just three published examples: namely a short segment from HIF1α bound to two partners, the TAZ1 domain and the asparagine hydroxylase FIH protein [64], a short segment from the C-terminus of p53 bound to four partners, S100ββ, sirtuin, CREB binding protein, and cyclin A2 [46], and a larger collection of various short segments bound to multiple partners [71].

Our decision to test whether hub proteins depend on disorder was motivated by prior experiments showing that conformational disorder enabled one particular protein region to bind to multiple partners [72]. We have carried out data mining on the Protein Data Bank (PDB) to find additional examples of both one-to-many and many-to-one complexes at atomic resolution.

We have found well over 300 sets that contain segments having the same sequence bound to two or more partners, but here we are focusing on unambiguously the

same protein bound to highly divergent partners (e.g. partner pairs with less than 25% sequence identity), thus reducing the numbers down to 23 sets of segments that bind to 2 to 9 partners. The goal is to provide detailed analyses of the conformational changes enabling the same disordered segment to bind to more than one protein partner. Overall these data support the view that the flexibility of disordered regions is a significant factor in the ability of IDPs to bind to two or more partners. As we assembled this dataset, we also found that alternative splicing events (ASEs) and PTMs were also involved in the process of enabling one disordered region to bind to more than one protein partner. These latter findings suggest that interplay of multiple factors has participated in the evolution of complex protein-protein interaction networks and might be important in the development of tissue-specific signaling networks.

Our data mining of PDB yielded over 500 sets that contain multiple, different MoRF segments bound to common binding partners, but here we are focusing on those larger domains (greater than 70 amino acids) bound to nonidentical MoRFs, thus reducing the number down to 160 sets of domains that bind to 2 to 48 segments. Our goal is to look at the detailed binding profiles of many-to-one binding and to perform structural analyses on the different binding segments. Two main binding profiles were observed in the assembled dataset. The MoRF segments sometimes bind to completely independent sites. Alternatively, the segments can bind to overlapping regions, which can range from highly similar sites to minimally intersecting sites on the corresponding partner. To quantitate the degree of overlap within the 5507 overlapping MoRF pairs in our 160 many-to-one set, we estimated the amount of spatial superposition each pair, which was expressed as a volume overlap ratio. This measure follows a normal

distribution when all the atoms of each MoRF are included. However, if only the backbone atoms are included or if the backbone atoms + C-beta atoms are included, then the distribution becomes much more asymmetric, showing steady numbers of pairs as the overlap ratio increases from very low overlap to almost 50% overlap, at which point the number of pairs increases rapidly. These results suggest that, in our dataset, similar binding sites for MoRF pairs are more common than are intersecting binding sites for MoRF pairs.

The detailed findings and results regarding the binding diversity and partner selection in protein disorder are described in the following chapters, thus leading to a better understanding of MoRF-domain network biology and regulatory mechanisms based on IDP regions. We expect that this improved understanding will eventually lead to deeper explanations of many cellular and biological processes. Hopefully, the specific examples we collected and analyzed in one-to-many, many-to-one and many-to-many binding mechanisms in this study will be seen to reveal the complexity and natural beauty of the protein interactome in cells.

CHAPTER 2

Materials and Methods

2.1. MoRF Datasets Preparation

Our disordered hub dataset was extracted from PDB by analyzing the complex structures that have short non-globular protein fragments bound to large globular structured partners. In this paper, we concentrated on those MoRFs which are short non-globular protein fragments whose visible residues in crystallographic electron density maps included between 5 and 25 residues and binding partners are globular proteins greater than 70 amino acids in length. The PDB entries we used were released on March 28, 2008 and June 19, 2012.

An interface size (ΔASA) of 400$Å^2$ was used to discriminate biologically relevant interactions and non-biological interactions caused by crystal packing contacts in this study [73]. The same cutoff was previously chosen by the authors of the protein quaternary structure file server (PQS), since the minimal ΔASA of homo-dimers and hetero-dimer are about 370 $Å^2$ and 640 $Å^2$, respectively [74].

2.2. Characterization of MoRF Clusters that Perform One-to-Many and Many-to-One Binding

Besides p53 other MoRFs that bind to two or more partners and that have structures in PDB have not been systematically compared to understand how disorder can bind to multiple partners. To discover specific disordered regions binding to multiple structured partners like p53, we used a Fasta program to align each MoRF sequence to the UniProt sequence database. This database encompasses the UniProtKB/Swiss-Prot

and UniProtKB/TrEMBL databases. The e-value was set at 1000 while carrying out the similarity search. Following that, we only kept those MoRFs which had overlapping regions (circled ones in Figure 6) in their parent sequence mapping and used a cluster algorithm (wherein at least one residue overlapped with the rest of the MoRFs in the same cluster).

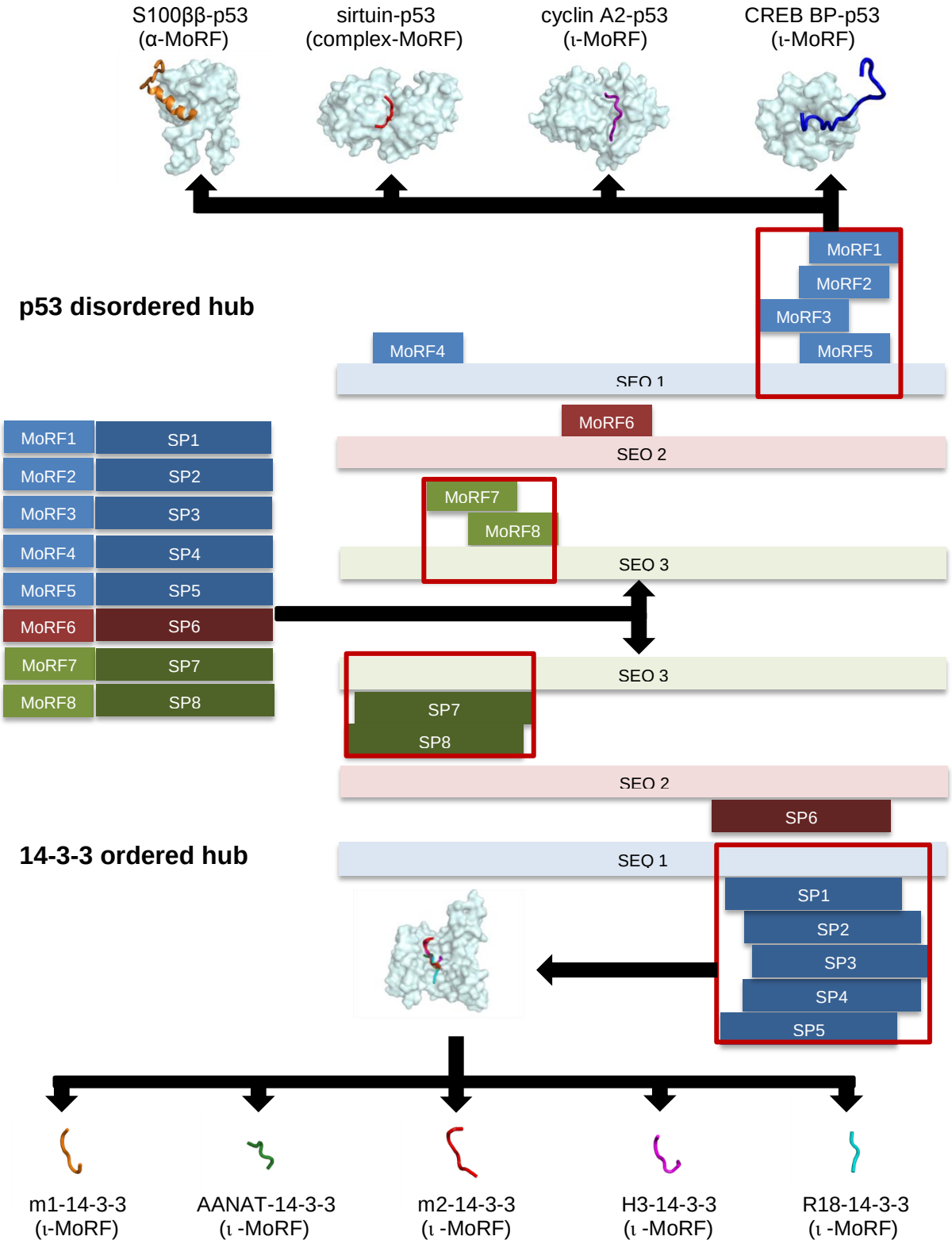


Figure 6. A schematic diagram to show how we constructed our (A) one-to-many and (B) many-to-one binding dataset by aligning and clustering MoRF sequences from complex structures in PDB.

2.3. Removal of Redundant MoRFs in MoRF Clusters

As our research is focused upon those MoRFs from the same disordered region which bind to structurally different partners, we used the blastcluster program to remove any redundant structured partners in our dataset based on 100% and 25% sequence identity. That means that those specific MoRFs are in one disordered region, but they use distinct residues to form bonding with different structured partners.

2.4. Removal of Atypical MoRFs in MoRF Clusters

After examination of the entire MoRF dataset manually, we found there were several unanticipated cases that were not consistent and needed to be removed from our dataset. They include the cases involving one MoRF interacting with more than one partner in a single PDB entry or a partner molecule which may be a subset of another partner in the same cluster.

2.5. Secondary Structure Assignment on MoRFs

We classified MoRFs into 4 different types (α, β, ι and complex) based on their secondary structure type which has the largest percentage value of the four types mentioned above. If there is no clear preponderance of any one secondary type (which is at least 1% greater than the other 2 types), we classified it as a complex-MoRF. Only the residues on the interface were counted. DSSP was used as the secondary structure assignment program here.

2.6. Sequence and Structure Similarity Analyses

The root mean square deviation (RMSD) of pairwise proteins was calculated by CEalign [75]. The coverage of alignable region is calculated by length of aligned regions dividing by average length of all sequences. The transposed coordinates and multiple

structure alignments were generated by MultiProt algorithm [76] using the complex structures including both MoRF and partner. Sequence identity calculations are based on the structure alignments. The sequence identities of MoRFs within many-to-one clusters were obtained from PRALINE multiple sequence alignment server [77]. The overlap ratio for each MoRF pair was calculated as the formula below, where V is the volume of the molecule. Vij means the union volume of MoRF i and MoRF j.

$$\text{Overlap ratio (OR)} = \frac{\text{Vi+Vj-Vij}}{\text{min(Vi,Vj)}}$$

Both residues in each aligned pair were compared to see if they are both in the binding or nonbinding region. The alignment will be considered identical only when the position in both proteins is assigned in the same class: either binding or nonbinding. For the case with more than 2 partners, we averaged all the identities together. Those aligned residues not consistent with their binding/nonbinding status (one is on binding region, but the other one is not) will be classified into another category that didn't show on our results. Here, those residues with higher solvent surface changes (greater than 1 $Å^2$) will be considered as interacting residues. Error bars that represent the 95% confidence interval (CI) of a mean are approximated from 3000 random samplings with replacement generated by the bootstrapping method. The molecular images in Figures were generated by PyMol software.

2.7. Peptide-Protein Interaction Annotation

Several immune-related protein interactions are considered as peptide-protein interaction. Interactions involving in MHC molecules, antibodies and T-cell receptors within our dataset are separated from other protein-protein interactions.

2.8. SCOP Classification on MoRF Partners

Structural Classification of Proteins (SCOP) is a database providing detailed and comprehensive annotations of the structural and evolutionary relationships between the proteins whose structure are known in PDB. The SCOP classification of proteins was constructed manually by visual inspection and structural comparison with assistance of tools. There are four levels existing in the SCOP hierarchy. Each protein can be assigned to reflect both structural and evolutionary relatedness.

1. Family: clear evolutionarily relationship (>30% pairwise sequence identity).
2. Superfamily: Probable common evolutionary origin (low sequence identity with structural and functional features suggesting a common evolutionary origin).
3. Fold: Major structural similarity (same major secondary structures in the same arrangement and topological connection).
4. Class: Types of folds, including all alpha, all beta, alpha and beta (a/b), alpha and beta (a+b), multi-domain proteins and so on.

SCOP 1.75 release (23 Feb 2009) was applied to our MoRF dataset on partner side to see if there is a structural preference for MoRF partner selection. There are 1195 folds, 1962 superfamilies, 3902 families, 38221 PDB entries and 110800 domains in the current release (excluding nucleic acids and theoretical models).

2.9. Network Analysis of MoRF Datasets

A summarized protein interaction network between the 510 human proteins in our MoRF set was generated by the Search Tool for the Retrieval of Interacting Genes/Proteins (STRING). STRING is a database of known and predicted protein interactions based on demonic context high-throughput experiments, conserved

coexpression and previous knowledge. The current STRING 9.05 database covers 5,214,234 proteins from 1133 organisms. The edges between MoRF nodes in the graph are based on the method of known and predicted interactions according to the following sources: neighborhood, gene fusion, co-occurrence, co-expression, experiments, databases, text mining, and homology. The MoRFs in the generated interaction network by STRING is highly connected which indicates MoRFs do perform functions appropriate for hubs.

CHAPTER 3

Binding Diversity of Intrinsic Protein Disorder

3.1. One-to-Many Binding

We identified 4289 MoRFs from the PDB based on their sequence length (5 to 25 residues). Of these, 452 complexes with small surface areas of interaction (<400 $Å^2$) were eliminated due to uncertainty regarding the biological significance of the interactions. An additional 689 complexes were excluded because their partners were nonglobular (length < 70 residues).

In order to identify overlapping MoRFs, MoRF sequences were mapped back to their parent sequences. A short segment will give exact matches to many unrelated sequences. Since many of the MoRFs are short, only 1805 of the remaining 3148 MoRFs could be unambiguously mapped in an automated fashion to their parent sequences in UniProt database. In addition, the parent sequence information are not always annotated in PDB. Based on the overlapping regions in parent sequence mapping (at least one residue), 298 MoRF sets with multiple partnerships were obtained. Structurally redundant partners were discarded from our final dataset based on imposing an upper bound of 25% pairwise sequence identity for every pair of partners.

Finally, 23 MoRF clusters with 61 partners were further confirmed by manual inspection to ensure that short peptides were bound to globular partners. Thus, for the dataset investigated herein, each MoRF associates with an average of 2 to 3 distinct partners. A summary of the development of the dataset is given in Table 2. Figure 7 is a bubble chart showing the 3-way relationship between MoRF length (x-axis), MoRF count

(y-axis) and cluster count (size of bubbles) in the 23 MoRF clusters. The 23 MoRF examples are listed in Table 3. The previous two partnerships involving HIF1α was not found in this study because the length of the peptide, 51 amino acids, exceeded the upper bound of 25 residues used in this study. Here, peptides are defined to have lengths in between 5 to 25 residues and domains are defined to have more than 70 (2008 MoRF dataset) or 40 (2012 MoRF dataset) residues. On the other hand, note that the previously described four partnerships involving the carboxy terminal tail of p53 were all found in our dataset [78], showing that our overall strategy found a previously known example the length of which was between our upper and lower thresholds.

Table 2. Description of one-to-many MoRF dataset.

Data set	MoRFs	Clusters	MoRFs per cluster
Initial MoRF dataset (5-25) [a]	4289		
MoRF dataset with biological interaction (>400Å^2) [b]	3837		
MoRF dataset with globular partner (>70) [c]	3148		
MoRFs mapped to UniProt sequence database [d]	1805		
MoRFs with overlapped region in mapping [e]	1493	298	5.01
MoRFs without 100% sequence identity in partners	248	87	2.85
MoRFs without 25% sequence identity in partners	214	77	2.78
MoRFs without atypical cases [f]	61	23	2.65

[a]MoRFs with 5 to 25 residues are the focus of this study.

[b]400 Å^2 cutoff was set to filter out the spurious interactions caused by crystal contacts.

[c]Binding partners of MoRF are supposed to be globular proteins having more than 70 residues to fold into a certain conformation. The excluded ones includes interactions between short domain like SH3, chromodomain, A/B chain of insulin , Gramicidin-form ion channels, peptides forming amyloid-like fibril, alpha-helical coiled coil, de novo proteins.

[d]Most MoRFs can't be mapped to UniProt are with 5 to 9 residues in length.

[e]MoRFs having one or more overlapping residues with each other.

[f]Atypical cases include, for example, one MoRF bound more than one partner the same PDB entry and partners with subsequences that exactly match the entire sequence of another partner.

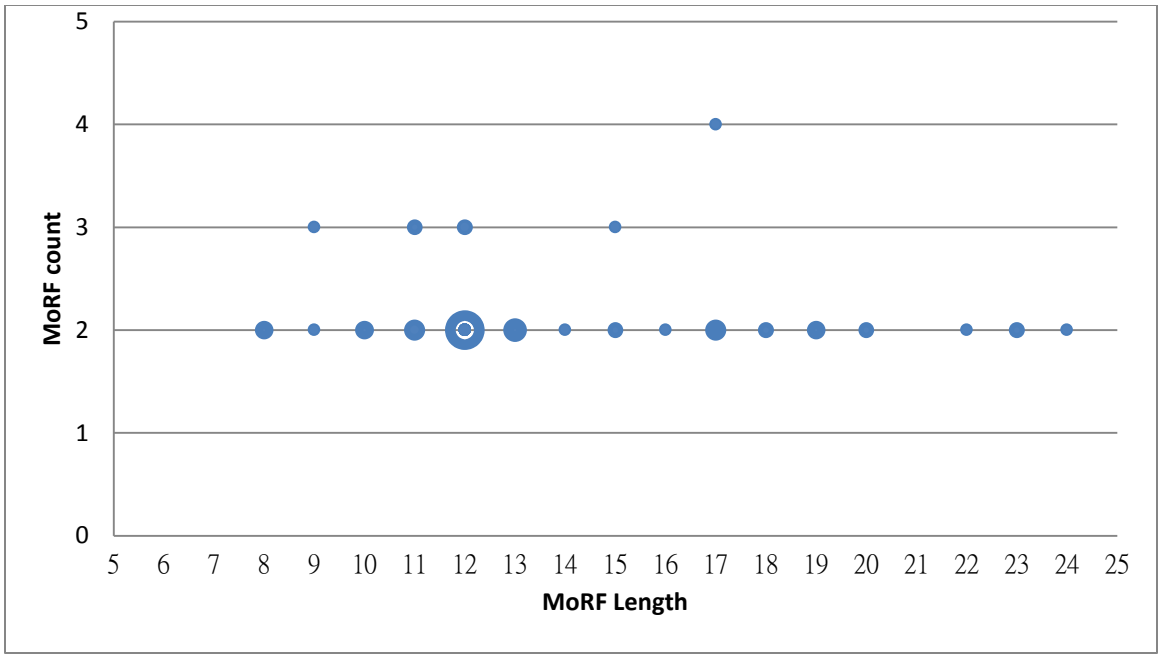


Figure 7. A bubble chart shows a 3-way relationship between MoRF length, MoRF count and cluster count in our one-to-many MoRF set.

Table 3. Twenty-three examples of MoRFs and their secondary structures.

MoRF examples		Bound conformation					Partners		MoRFs	
		N	Helix	Sheet	Coil	Complex	RMSD	Coverage	PTM	AS
8 MoRFs with differently-folded partners		26							11	1
1.	Histone H3 – N-term I	9	0	0	9	0	7.07	0.21	5	-
2.	p53 – near C-term	4	1	0	2	1	6.80	0.39	1	[1]N
3.	CTD of RNA polymerase II	3	0	0	3	0	8.35	0.26	3	-
4.	Angiotensin	2	0	0	2	0	7.74	0.27	0	-
5.	HIV envelope glycoprotein	2	0	0	2	0	4.16	0.41	0	-
6.	Histone H3 - N-term II	2	0	0	2	0	8.25	0.22	2	-
7.	Vasopressin	2	0	0	2	0	8.69	0.37	0	-
8.	p53 – near N-term	2	2	0	0	0	6.18	0.62[#]	0	[1]Y
15 MoRFs with similarly-folded partners		35							2	4
9.	Nuclear receptor coactivator 1 & 2	5	2	0	2	1	3.94	0.92	0	-
10.	Nuclear receptor corepressor 2	3	2	0	1	0	3.43	0.85	0	[1]TS
11.	TRAP 220	3	3	0	0	0	3.05	0.91	0	[2]Y
12.	Nuclear receptor coactivator 1	2	2	0	0	0	5.49	0.85	0	-
13.	BAK peptide	2	2	0	0	0	5.50	0.73	0	[1]N
14.	Nuclear receptor 0B2 – N-term	2	2	0	0	0	3.74	0.86	0	[1]N
15.	Troponin I, cardiac muscles	2	0	0	1	1	3.01	0.79	0	-
16.	Nuclear receptor 0B2 – C-term	2	1	0	1	0	3.88	0.80	0	[1]N
17.	Cell death protein GRIM	2	0	2	0	0	2.33	0.79	0	-
18.	Beclin-1	2	2	0	0	0	4.10	0.84	0	-
19.	Histone H4	2	0	0	2	0	3.93	0.50*	0	-
20.	Bcl-2-like protein 11 (Bim)	2	2	0	0	0	2.72	0.90	0	[1]Y
21.	Amyloid beta A4 protein	2	0	0	2	0	2.93	0.84	0	[1]Y
22.	Rhodopsin	2	2	0	0	0	4.25	0.86	0	-
23.	DNA repair protein RAD9	2	0	0	2	0	3.53	0.36*	2	-

N: numbers of MoRFs in the set; PTM: post-translation modification; AS: alternative splicing; TS: tissue-specific alternative splicing. [#]Although most residues within the two partners can be roughly aligned together, their individual structure varies a lot. *Within these two sets, the coverage of good alignments is low because one partner is a sub-domain of the other partner but with low sequence identity. [1]MoRFs are from human; [2]MoRFs are from mouse; -MoRFs are from other species. Underlined MoRFs are in 11 sets.

Most sets contain one MoRF interacting individually with two partners, but six of the sets have more than two partners. These are the N-terminus of histone H3, nuclear receptor coactivator 1 and 2, the C-terminus of p53, the NR corepressor 2, the thyroid receptor associated protein 220, and the carboxyl-terminal domain (CTD) of RNA polymerase II. Since MoRFs in the NR coactivator 1 and 2 share similar sequences and can be mapped to the same parent sequence, our method clustered them together as a single set. Most clusters have MoRFs with similar secondary structures in different complexes. Only five of them exhibit a mixture of different secondary structures (Table 3).

The goal here was to find the same MoRF sequence bound to structurally distinct partners, so partners having low sequence identity were chosen. A sequence identity of 25% was chosen as the upper bound because proteins with sequence identities higher than this value are almost always similar in structure [79]. Nevertheless, even though the partners of each MoRF set were selected to have low sequence identity, several partner conformations turned out to exhibit structural similarity. Based on the structure alignment of their partners, the 23 MoRF sets can roughly be grouped into 15 MoRFs with similarly-folded partners (with ~19% sequence identity on average) and 8 MoRFs with differently-folded partners (with ~10% sequence identity on average). Notice that MoRFs with differently-folded partners apparently prefer to form irregular secondary structure upon binding, while MoRFs with similarly folded but sequence diverse partners tend to prefer to form helix or sheet (Table 4).

Table 4. The combination of secondary structure types in the 23 MoRFs.

Secondary structure	Clusters	Similarly-folded partners	Differently-folded partners
α + β + ι +Complex	0	0	0
α + β + ι	0	0	0
α + β + Complex	0	0	0
α + ι + Complex	2	1	1
β + ι + Complex	0	0	0
α + β	0	0	0
α + ι	2	2	0
α + Complex	0	0	0
β + ι	0	0	0
β + Complex	0	0	0
ι + Complex	1	1	0
α	8	7	1
β	1	1	0
ι	9	3	6
Complex	0	0	0

Two predictors, ANCHOR [47] and MoRFpred [48], have been developed to predict partner binding sites within longer regions of disorder. Application of these predictors to the MoRF-containing sequences shows that, while both predictors typically indicate binding sites corresponding to the observed MoRFs, neither predictor is particularly accurate with respect to the locations of the binding sites (data not shown). Interestingly, the locations of the MoRFs with similarly-folded partners are predicted with slightly greater accuracy by both predictors as compared to the locations of MoRFs that bind to differently-folded partners.

3.1.1. Fifteen MoRF Sets with Similarly-Folded Partners

Among the 15 MoRFs with partners having similar folds, similar binding profiles and common interacting residues were observed. Partner pairs within 11 of these MoRFs have both a relatively low RMSD and a relatively good structural alignment. The RMSD values and the fraction of the total residues that gave good structural alignments were estimated for the sets of partners for the 18 MoRFs. The 11 sets that had both a relatively low RMSD (2.33 to 5.49 Å) and a relatively high fraction with good structural alignment (termed coverage; values between 0.79-0.91), are given in Table 5. The partners of the DNA repair protein RAD9 have a reasonable RMSD (3.53) but a low coverage (0.36). This protein was not discarded because one of the partners had a large, non-alignable extra domain that was responsible for the low coverage.

In order to answer if having similar binding patterns mean MoRFs tend to bind a specific set of residues on partners, we analyzed the sequence identities of binding regions and nonbinding regions on partner side to determine whether the interacting residues are more selected during evolution (Table 5). Those interacting residues tend to

be selected to form connection with same MoRF using similar binding patterns as we expected. The sequence conservation of the binding region is significantly higher than those in other parts of the protein. The mean sequence identity for structurally aligned binding and non-binding residues are 42±6% and 20±3%, respectively, within these 11 sets. Binding residues, which are usually on the surface, have about 2.5 fold higher sequence identity than nonbinding surface residues, indicating that these interactions are likely to be biological significant. These averages were taken over structurally matching residues. In comparison, for a collection of enzymes with ~ 25% sequence identity, the active site residues exhibit sequence identities in the range of 43% to 70% [80]. Interestingly, the binding residues being discussed here for several of the proteins show sequence identities that overlap the lower part of the observed range for enzyme active site residues, which are known to have a high tendency to be conserved. Note also that the binding residues show a much higher conservation for the aligned residues as compared to the nonbinding surface residues (column B versus NB_E) or to the nonbinding buried residues (column B versus NB_B).

For the same MoRF bound to structurally similar partners, only slight conformational changes of MoRF side chains were observed, whereas the backbone conformations of the same MoRF between various complexes are relatively uniform. Many features of the various interactions indicate that the disordered binding segments were likely to have been disordered before binding. Hence, these results also add further weight to the existence and function of intrinsically disordered regions inside cells.

Table 5. Sequence identities of binding regions and nonbinding regions in partners of 11 disordered hub examples based on MultiProt structural alignment.

Disordered hub examples	B	NB	NB_B	NB_E	Overall
Nuclear receptor corepressor 2	0.36	0.17	0.22	0.15	0.21
TRAP 220	0.51	0.23	0.26	0.19	0.24
Nuclear receptor coactivator 1	0.62	0.16	0.22	0.08	0.16
Nuclear receptor 0B2 – near N-term	0.47	0.20	0.30	0.16	0.19
Troponin I, cardiac muscles	0.36	0.34	0.62	0.29	0.25
Nuclear receptor 0B2 – near C-term	0.27	0.16	0.13	0.25	0.12
Cell death protein GRIM	0.33	0.28	0.60	0.09	0.26
Beclin-1	0.45	0.17	0.26	0.06	0.17
BCL-2-like protein 11	0.44	0.16	0.26	0.11	0.21
Alzheimer's disease amyloid A4 protein	0.33	0.19	0.33	0.11	0.15
DNA repair protein RAD9	0.43	0.27	0.33	0.11	0.07
	0.42±0.06	0.20±0.03	0.26±0.05	0.16±0.03	0.19±0.04

The values in columns labeled B, NB, NB_B and NB_E give the averages of pairwise sequence identities of binding, nonbinding, nonbinding buried, nonbinding exposed residues of the specific MoRF partners based on structure alignments, respectively. The value in Overall column gives the sequence identity of all residues based on sequence alignment.

It is fascinating that 5 of 11 disordered hub examples belonged to the family of coregulatory proteins of nuclear receptor (NR), including thyroid receptor associated protein 220 (TRAP 220). Our dataset indicates that protein disorder is involved in coregulatory proteins of nuclear receptors (NR) such as coactivators and corepressors, making it possible for them to perform one-to-many signaling and to function as disordered hubs. The nuclear receptors (NRs) are a super-family of proteins, associated with other coregulatory proteins involved in the direct mediation and control of the expression of specific gene transcription in response to sensing the presence of hormones and other molecules. Recent data shows that, in addition to direct activation of the basal transcription machinery, nuclear receptors inhibit or enhance transcription by attracting an array of coactivator or corepressor proteins to the transcription complex.

NRs may be classified into two broad categories according to their mechanism of action and subcellular distribution in the absence of ligands. Ligands bind to type I NRs in the cytosol resulting in the dissociation of heat shock proteins, the formation of homo-dimers, translocation from cytoplasm into the cell nucleus, and binding to hormone response elements. Type I NRs include NR subfamily 3, which encompass androgen receptors, estrogen receptors, glucocorticoid receptors and progesterone receptors. Type II NRs, in contrast to type I NRs, are retained in the nucleus and heterodimerize upon binding to DNA in the absence of a ligand, when type II NRs are usually bound to a corepressor. Ligand binding to type II NRs triggers the dissociation of the corepressor, leading to the initiation of transcription by the coactivator. Type II NRs include NR subfamily 1, and receptor molecules such as retinoic acid receptor (RAR), retinoid x receptor (RXR), thyroid hormone receptor (TR) and vitamin D receptor (VDR).

Peroxisome proliferator-activated receptor-binding protein (PPAR or PBP), also known as thyroid hormone receptor-associated protein 220 (TRAP 220), is an anchor for multi-subunit mediator transcription complex. It functions as a transcription coactivator for nuclear receptors. These coactivator proteins often exhibit histone acetyltransferase (HAT) activity, which weakens the association of the histone to DNA, therefore promoting gene transcription. Three MoRFs in PBP have been found to be involved in multiple interactions with various type II nuclear receptors, such as vitamin D3 receptor (VDR), retinoic acid receptor-beta (RAR-beta) and retinoid x receptor-alpha (RXR-alpha) in our dataset. RAR or VDR, when forming a heterodimer with RXR, can bind to hormone response elements, forming a complex with corepressor protein in the absence of any ligands. When a ligand acting as agonist binds to RAR or VDR, it results in the dissociation of the corepressor and recruitment of coactivator which in turn, promotes transcription of downstream target gene.

When gene transcription is repressed by nuclear receptors, it is mediated by interactions with corepressor proteins. This reaction, in turn attracts histone deacetylases (HDACs) to the chromatin, triggering the strong binding of histone to DNA; thus repressing gene transcription. The antagonist further reinforces the binding of corepressor to the nuclear receptor. MoRF mechanisms are also involved in the down regulation of target gene expression when the nuclear receptor corepressor 2 binds to related nuclear receptors such as peroxisome proliferator activated receptor (PPAR), estrogen related receptor gamma (ERR-gamma) and progesterone receptor (PR) [81-83].

In the previous two examples, such coregulatory proteins can interact with various receptors with low sequence identity but high structure similarity using the same MoRF

region. The configurations of secondary structures in those MoRFs bound to receptors are also comparable. Further experiments and analyses provided us more detailed and specific explanations regarding how disordered regions facilitate the binding diversity in different complex structures. The analyses of solvent surface area profiles from the examples with structurally different partners, shows different interfaces accommodate a variety of binding partners and those overlapping residues in interfaces bind to different molecules to varied extents. Otherwise, analogous binding profiles were observed within our 11 examples with similar partnerships.

Three examples were chosen from our 11 sets in order to assess in more detail how a disordered region uses its conformational flexibility to form interactions with similar but not identical binding pockets. The three examples can be described as an α-MoRF, a β-MoRF and an irregular-MoRF corresponding to the thyroid receptor associated protein 220 (TRAP220), the cell death protein GRIM and the Alzheimer's disease amyloid A4 protein homolog, respectively. Figure 8 shows the interacting residues and binding sites of the three selected cases.

Figure 8.A shows four important residues (M5, L6, L9 and L10) on TRAP220 stretching into the clefts on the surfaces of receptor proteins with small structural variations. On the contrary, those residues on the non-buried side of the α-MoRF have larger conformation fluctuations over the three complexes.

Alanine 1 and tyrosine 4 contribute most of the buried surface areas to the interaction of GRIM and IAP1 (Figure 8.B). The tyrosine side chain makes a huge rotation to fit distinct cavities of IAP1 while the backbone's β-sheet conformation and key interactions didn't change too much (Figure 8.B.1 and 8.B.2).

The MoRF in amyloid A4 protein is mostly coil but has local regions that could be classified as sheet and helix according the DSSP. Its central region (NPTY) adopts coiled conformation and maintains comparably similar structure (colored in black) in both complexes (Figure 8.C). The N-terminus (NGYE) of the green MoRF (Figure 8.C.1) stays in a coiled structure while the N-terminus of the red MoRF in (Figure 8.C.2) turns into β-strand to form an anti-parallel β-sheet with another strand on DAB1 protein. The spatial arrangement of a tyrosine 4 was observed to change substantially, suggesting that this change may facilitate the binding to two different surfaces by the same sequence.

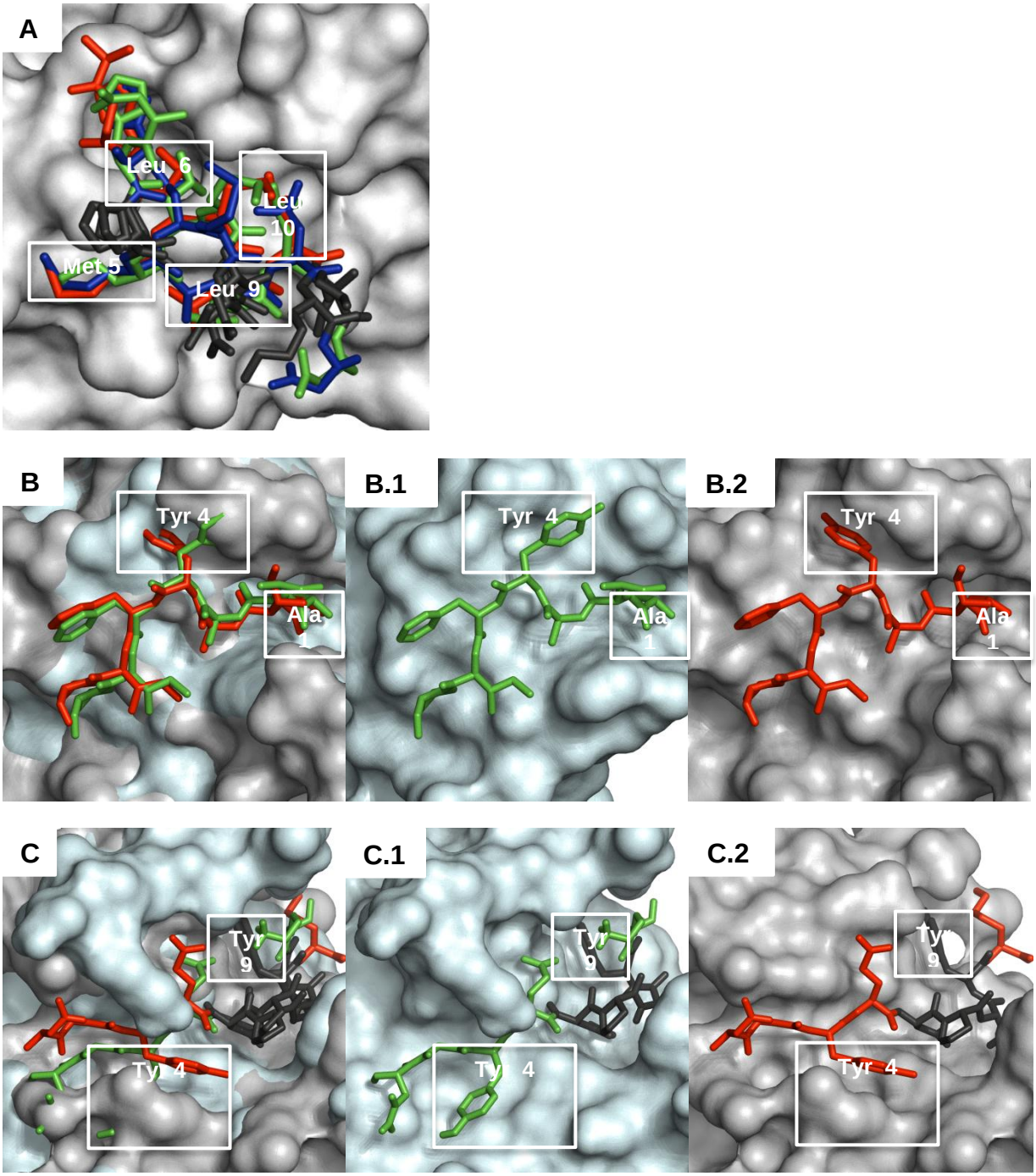
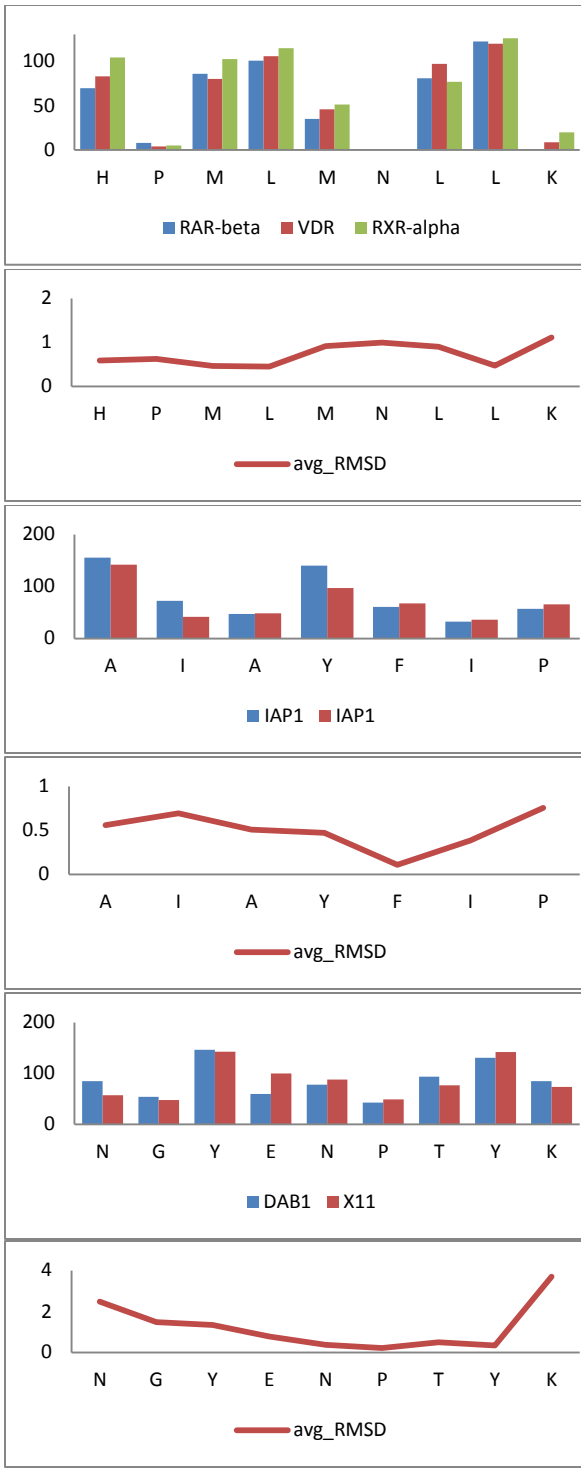


Figure 8. Conformational changes and variations of an identical MoRF binding to its structure-homologous partners from the three selected examples. Different MoRFs are shown in different colors. (A) A fragment from TRAP220 forming α-MoRFs to associate with VDR, RAR-beta and RXR-alpha. Those residues on the exposed side of helixes are colored in black. (B) The binding sites of a β-MoRF from GRIM and apoptosis 1 inhibitor (IAP1). (C) The irregular-MoRF in amyloid A4 protein adapts a highly flexible structure to accommodate the binding pockets of DAB1 and X11. Four structurally constrained residues with lowest RMSD are shown in black.

In addition to gathering general evolutionary information for the whole interfaces that MoRFs associate with, some further calculations were carried out on the same three selected examples in order to explore more details at the residue level. We compared the partner residues with which each MoRF residue associates to determine if partner residue variability correlates with overall conformational variation of MoRF itself (Figure 9.A, 9.B and 9.C). Our hypothesis was: the more diverse the amino acids with which a MoRF residue associates, the greater the structural variability of the MoRF backbone. However, the correlation analyses between diversity of partner residues (not shown in Figure 9) and averaged root mean square standard deviation (RMSD) on C-alpha atom (line plots in Figure 9) did not show an obvious and strong correlation. While our particular hypothesis was not supported, Figure 9 nevertheless contains interesting summary information regarding the changes that are observed when one MoRF binds to multiple partners.

Note alternating burial and nonburial of residues (Figure 9.A); this pattern can be traced to the α-helical structure (e.g. an α-MoRF) of the thyroid receptor associated protein 220. Also notice that the buried residues are more hydrophobic, and, except for proline, the nonburied residues are more hydrophilic. While hydrophobic, proline is often found on the surfaces of proteins and, furthermore, frequently occupies positions near the amino-terminal ends of helices. Based on our data, prolines compare richly in MoRFs and other disordered regions with ordered regions; however, there is no significant difference between MoRFs and other disordered regions. We speculate that proline serves as a structure breaker and disorder-promoting residue, playing an important character to maintain MoRF regions' flexibility until the binding events.



A

		1rkh_A	1xdk_B	1xdk_A
K	1	K		
N	2	E		VEP
H	3	IKEV	IKE	VREEP
P	4	E	E	E
M	5	IPLE	PLE	FVTFE
L	6	IILKLEV	VILLEM	VVLRFEM
M	7	SIE	IIE	LDVE
N	8	K	K	VV
L	9	QIKL	IVKLE	FVEKF
L	10	IKSQIL	VKFIQIL	VKFLQVL
K	11	K	K	VKL
D	12		K	K

B

		1jd5_A	1seo_A
A	1	LMDWNDQWE	IGSWEDERW
I	2	GLMDQWES	EIGSEDEW
A	3	GLMWL	VEIGW
Y	4	TRRGLM	KEIGS
F	5	RCGGGL	KCGVE
I	6	GGWL	VEW
P	7	CGL	CGVWS
D	8		

C

		1oqn_B	1x11_A
Q	1		MQDIG
N	2	DYIAKIR	MYIADQ
G	3	DYIAKR	KMYIAD
Y	4	SYIRFQ	KSYIADAQGF
E	5	RSYIF	TKRSYIF
N	6	VHISYF	LRISYFY
P	7	FQY	SFSY
T	8	VHIFIYK	LRFAYF
Y	9	ARVHEISGEH	KRTISRKQMY
K	10	HGKE	TDGKY
F	11		GKYFL
F	12		RRKRF

Figure 9. Binding surface area profiles (bar graphs), averaged RMSD (line plots), and important interacting residues with various globular partners (right) calculated on each residue from the three selected MoRF sets: (A) α-MoRF: thyroid receptor associated protein 220, (B) β- MoRF: Cell death protein GRIM and (C) irregular-MoRF: Alzheimer's disease amyloid A4 protein homolog. Notice in group (C) the specific interactions between 2 aromatic groups that oriented themselves orthogonally.

Several interesting points were found from the sequence and structure analyses of the two residues with the greatest contributions of solvent surface area in the GRIM-IAP1 interaction (Figure 9.B). Although the side chain of alanine 1 is relatively small, a fairly large area becomes buried into the interface. More detailed analysis shows that, for this residue, not only the side chain participates the formation of interaction, but backbone atoms also play a significant role.

Big rotations on side chains also related to higher backbone structure variations, such as the cases on L9 in the α-MoRF (Figure 9.A), Y4 in β-MoRF (Figure 9.B) and Y4 in irregular-MoRF (Figure 9.C). The low RMSD of P7 in the irregular-MoRF example (Figure 9.C) may correlate with its capping function in the edge of helix in the X11 protein.

Figures 10 and 11 show another two examples for which the flexibility needed to accommodate different partner surface features is manifested as side chain rotations. Lysine in nuclear receptor corepressor 2 has different conformations to stretch into the opposite cleft in three complexes to form the associations between the three receptors (Figure 10). Histidine and arginine in nuclear receptor coactivator 1 (NCOA1) and 2 (NCOA2) also act in a similar way in Figure 11. Here, the two different proteins NCOA 1 and 2 are grouped into one cluster in our dataset because both of them have similar conserved binding sequences containing LxxLL motifs (“HKILHRLLQD” and “HKILHRLLQE”) like other NR-boxes [84]. The side chain conformations of the three leucine residues stay nearly the same except for the ones that interact with the androgen receptor.

This example demonstrates that the same proteins can be involved in both one-to-many and also many-to-one binding, thus raising the level of network complexity and leading to multi-protein regulatory complexes that can respond to environmental signals. Comparing our one-to-many dataset described herein with our many-to-one dataset (will discuss in 3.2) reveals that, of the 23 examples in Table 3, there are 12 cases of proteins involved in both one-to-many and many-to-one binding. That is, 12 of the MoRFs in Table 3 bind to a structured partner that also binds to additional MoRFs having different sequences. Since our identification of one-to-many and many-to-one examples did not involve any steps for identifying MoRFs involved in both mechanisms, we find this number of 12 of 23 involved in both mechanisms to be quite high and to suggest that such dual use of both mechanisms is likely to be a very common feature of protein-protein interaction networks.

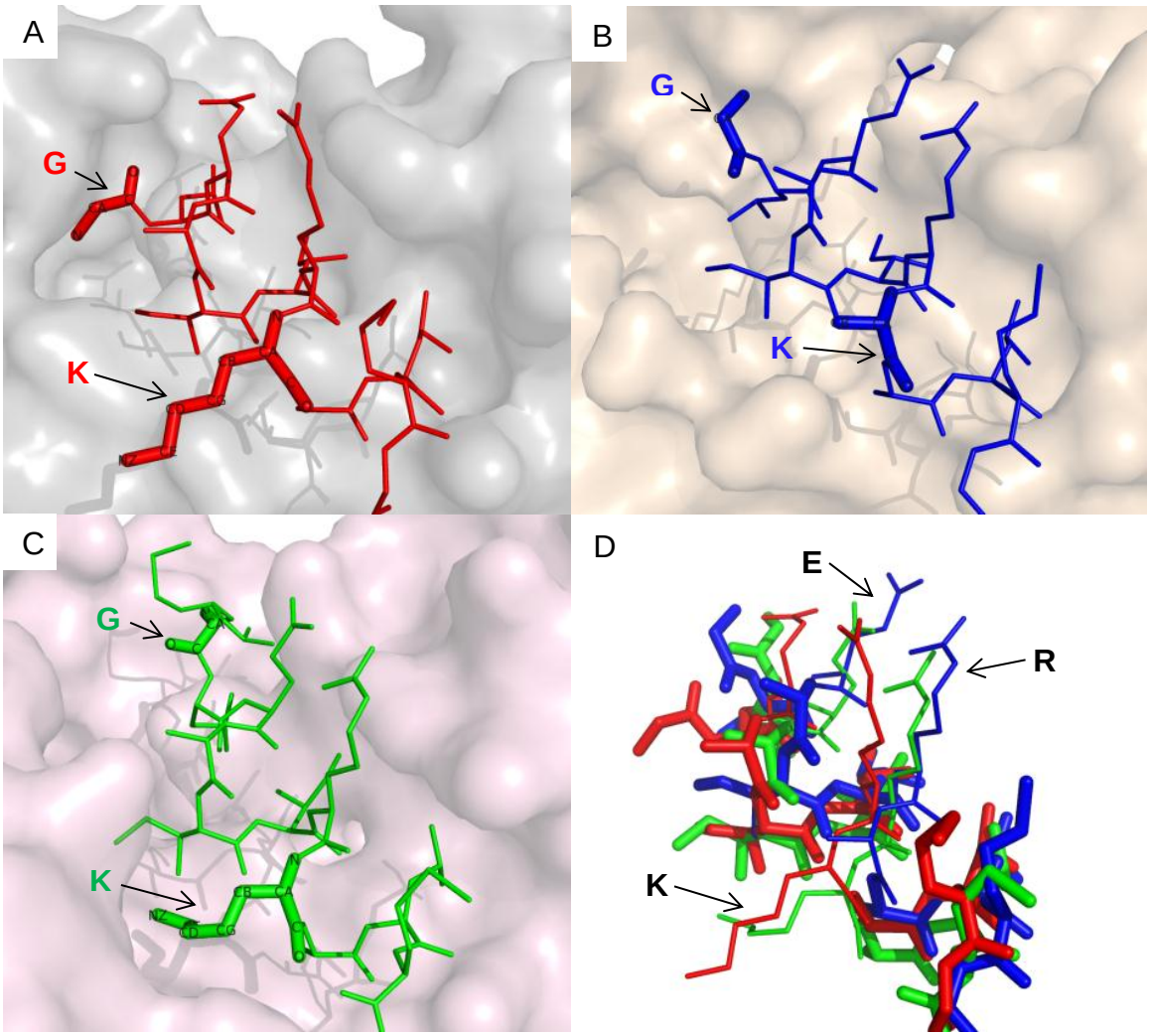


Figure 10. MoRFs in nuclear receptor corepressor 2 bind to 3 different but structurally similar nuclear receptors. (A) Estrogen-related receptor gamma (with α-MoRF in 2GPV), (B) Progesterone receptor (with α-MoRF in 2OVH), (C) Peroxisome proliferator activated receptor (with ι-MoRF in 1KKQ) and (D) The charged residues of the core MoRF region rotate more in the superimposition of the 3 complexes.
An interactive view:
http://imolecules3d.wiley.com:8080/imolecules3d/review/7RE6UdLGVvApooD5ECvQCKkGKHJKP7qdI7ZN22baIqoUGGY2LAR911FxPS6QN17b689/1288

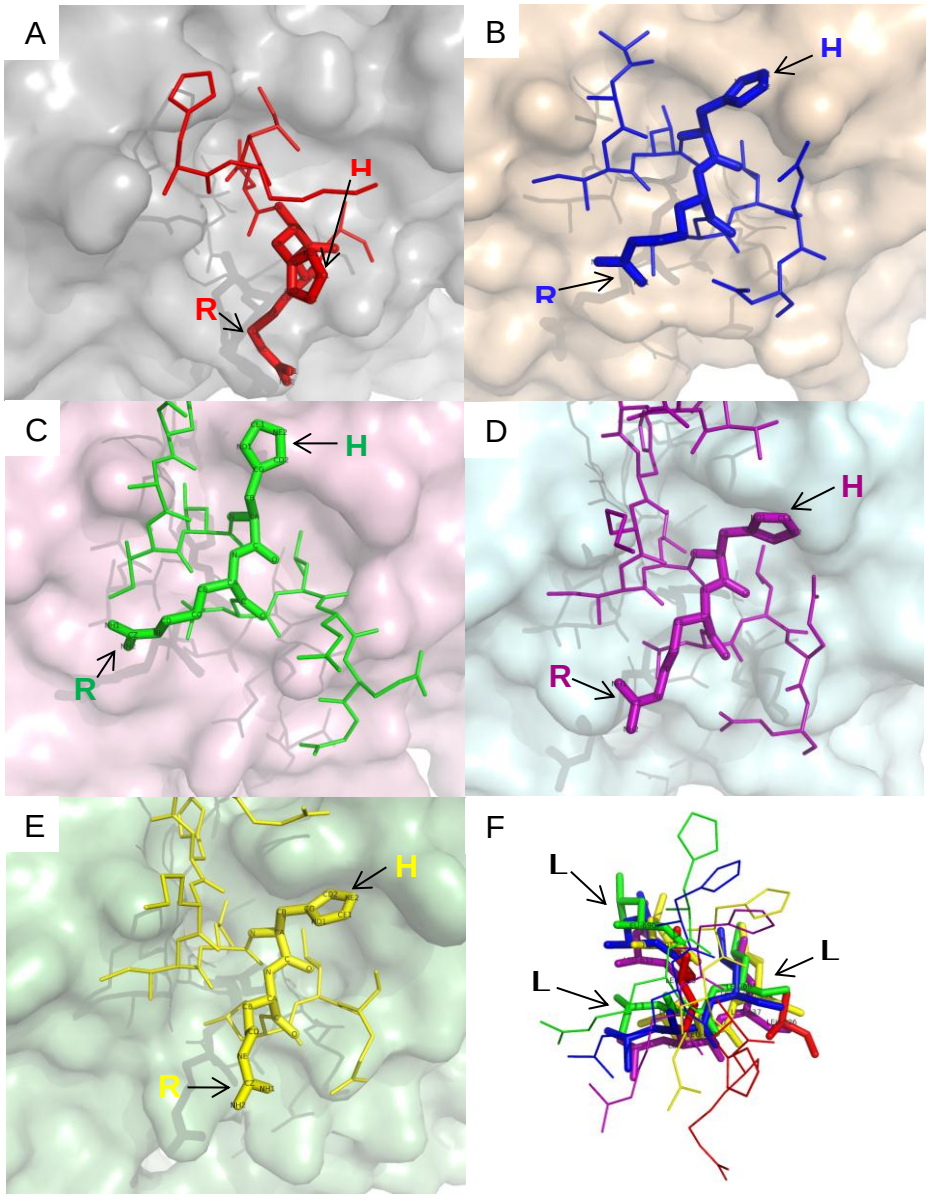


Figure 11. The diagram shows a variety of interactions of MoRFs with highly similar sequences in nuclear receptor coactivator 1 and nuclear receptor coactivator 2. (A) ι-MoRF in nuclear receptor coactivator 2 interacts with androgen receptor (1T65). (B) α-MoRF in Glucocorticoid receptor-interacting protein1 (alternative name of NCOA2) interacts with estrogen receptor (1L2I). (C) complex-MoRF in Nuclear receptor coactivator 1 isoform 1 interacts with orphan nuclear receptor NR1I3 (1XV9). (D) α-MoRF in nuclear receptor coactivator 1 interacts with bile acid receptor (2O9I). (E) ι-MoRF in nuclear receptor coactivator 1 isoform 3 interacts with orphan nuclear receptor PXR (3BEJ). (F) The 3 Leucine residues of the LxxLL motif are superimposed well in the 5 complexes.
An interactive view:
http://imolecules3d.wiley.com:8080/imolecules3d/review/7RE6UdLGVvApooD5ECvQCKkGKHJKP7qdI7ZN22baIqoUGGY2LAR911FxPS6QN17b689/1289

In summary, NCOA binding molecules include many kinds of nuclear receptors, including androgen receptor (AR), estrogen receptor (ER), nuclear receptor subfamily 1, group I member 3 (NR1I3/CAR), bile acid receptor (BAR) and pregnane X receptor (PXR) [85-89].

3.1.2. Eight MoRF Sets with Differently-Folded Partners

Eight MoRFs in our dataset converted into significantly different conformations to fit onto the surfaces of structurally different molecular partners. For these examples, only a small portion of their partners' residues can be structurally aligned. We selected the three examples with the largest number of partnerships (p53, RNA polymerase II (RNAP II) and histone H3) to illustrate the variable buried surface area of each MoRF residue upon diverse binding (Figure 12).

Charged residues (R, H and K), aromatic residues (F and Y) and phosphorylation-related residues (S, T, Y, H, R, K) in MoRF regions vary substantially in their contributions to binding different partners. In contrast, proline contributions to the different interfaces involving RNAP II remain relatively stable. Unlike MoRFs with similarly folded partners, which generally use their various residues in quite similar ways to associate with relatively conserved interacting residues, each partnership within this set utilizes conformationally distinct MoRFs and different residues or the same residues with different degrees of burial in their associations with their very distinct partners. That is, the same MoRFs show large variability in their side chain burial and exposure and even shifts in the binding region when binding to structurally divergent partners.

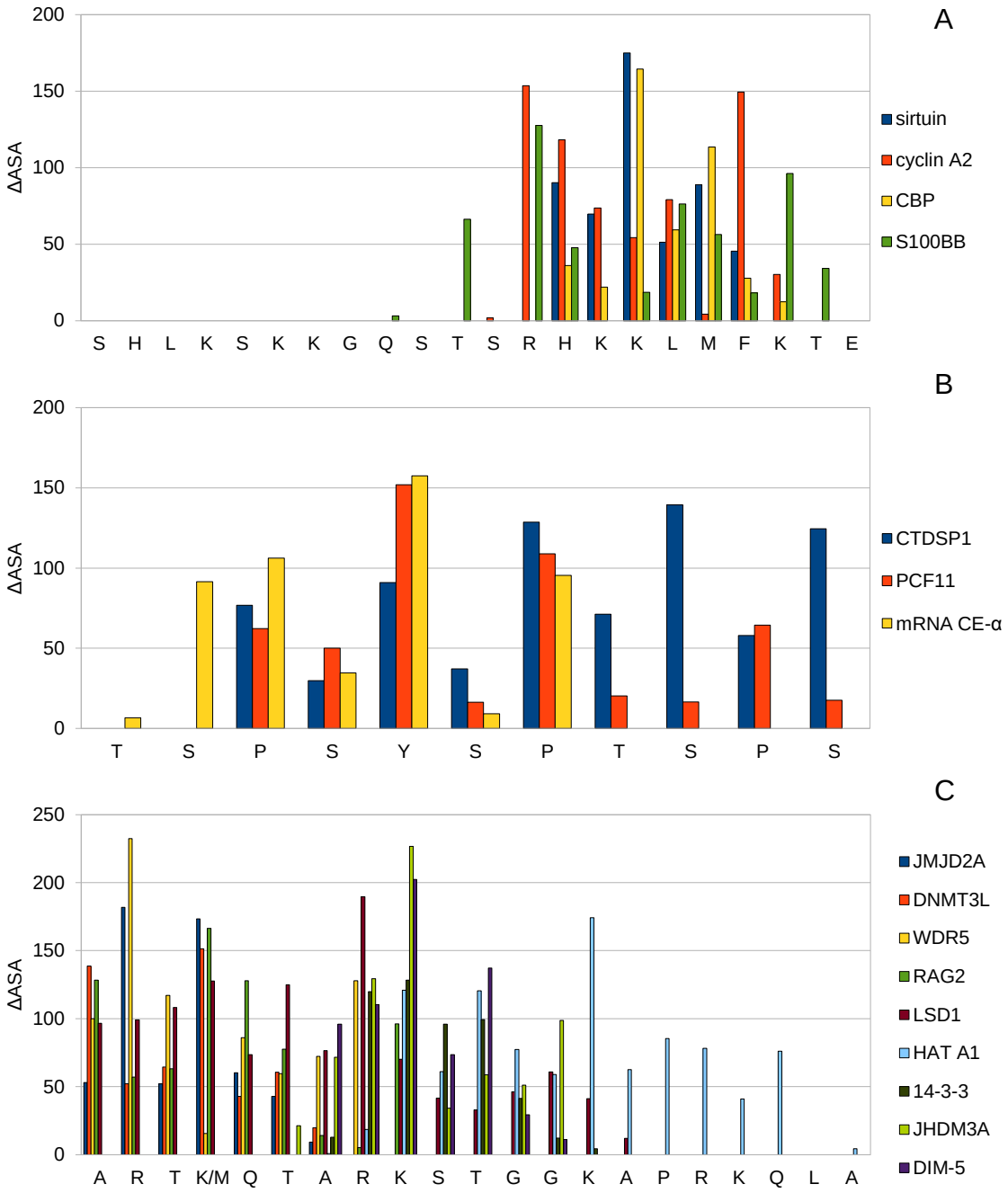
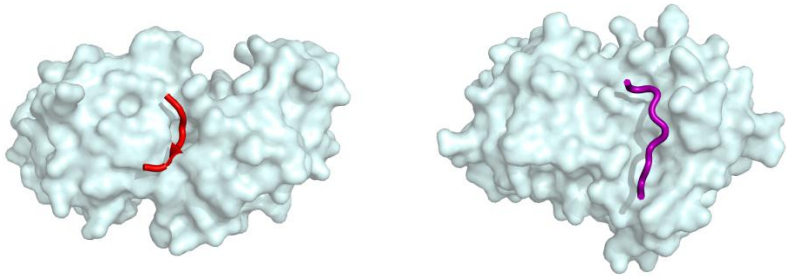


Figure 12. The profiles of solvent surface area changes within 3 selected MoRF clusters with structurally different partners: (A) p53, (B) RNAP II and (C) H3. The Y axis gives the change in surface area of each entire residue upon binding, while the X axis gives the residues.

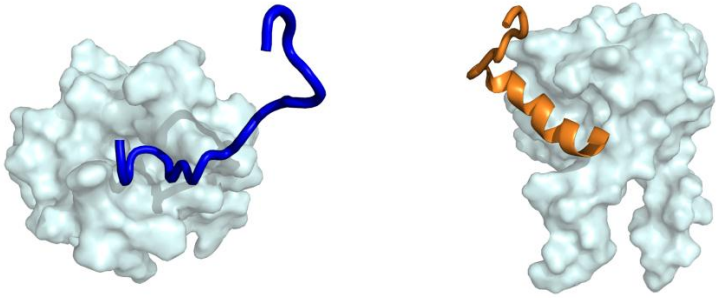
In addition to differential side chain burial and rotations, PTMs are also observed to be associated with the conformational alterations that are observed when the same MoRF binds to different partners, especially for those MoRFs that bind to structurally distinct partners. That is, of the 26 complexes involving differently-folded partners, 11 have posttranslationally modified residues. On the other hand, for the MoRFs with similarly folded partners, just 2 of the 35 complexes contain PTMs.

The C-terminus of p53 illustrates the conformational changes of a single MoRF within different partnerships. It was observed to transform either into a complex MoRF, an ι-MoRF (irregular MoRF), or an α-MoRF (helix), in four different structures in our dataset (Figure 13). The complex MoRF is composed of 3 residues of β-strand and 3 residues of coil and was classified as a β-MoRF in our previous work [78]. This change from the previous work arose because here we use automated secondary structure assignment (DSSP), whereas the previous work used the crystallographer's assignment of secondary structure.



A. Sirtuin & p53 B. Cyclin A2 & p53

A. Sirtuin																H	K	K	L	M	F					
B. Cyclin A2															S	R	H	K	K	L	M	F	K			
C. CREB BP					S	H	L	K	S	K	K	G	Q	S	T	S	R	H	K	K	L	M	F	K		
D. S100ββ					S	H	L	K	S	K	K	G	Q	S	T	S	R	H	K	K	L	M	F	K	T	E
p53/1-393	366	S	S	H	L	K	S	K	K	G	Q	S	T	S	R	H	K	K	L	M	F	K	T	E	G	389

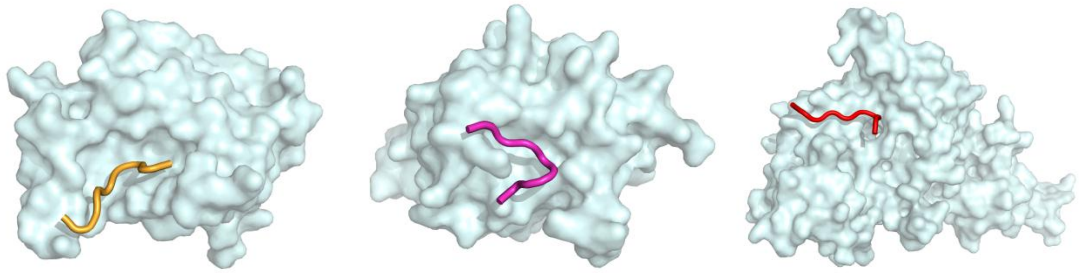


C. CREB BP & p53 D. S100ββ & p53

Figure 13. Four different biological molecules interact with C-terminus of p53. (A) Sirtuin: an NAD-dependent deacetylase (with complex-MoRF), (B) Cyclin A2 (with ι-MoRF), (C) CREB binding protein (with ι-MoRF), and (D) S100 calcium-binding protein (α-MoRF). In the sequence alignments, a residue having a posttranslational modification in PDB is indicated in red.

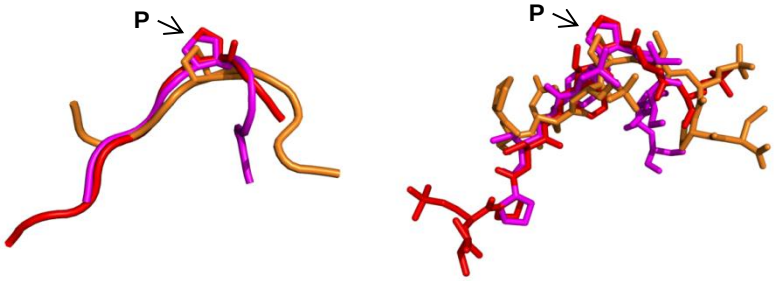
Although the two other MoRFs, RNAP II and H3, with distinctly folded partners have coiled structures for all of their 3 and 9 complexes, respectively, the backbone conformations differ markedly between any two pairs of structure.

The MoRF in the carboxyl-terminal domain (CTD) of RNAP II facilitates several enzymes to make post-transcriptional and post-translational modification by binding (marked in red in Figure 14). The CTD in RNAP II is composed of up to 52 heptapeptide repeats (YSPTSPS) which are important for polymerase activity [90]. Efficient capping, splicing and polyadenylation of mRNAs all require the CTD portion of RNAP II. For example, the CTD small phosphatase 1 (CTDSP1) catalyzes the dephosphorylation of Ser 5 within the tandem 7 residues repeats, causing the initiation of RNA polymerase II transcription (Figure 14.A) [91]. The Ser 2-phosphorylated CTD binds to a CTD-interacting domain (CID) in protein 1 of cleavage and polyadenylation factor I (PCF11), which is essential for transcription elongation 3' and RNA processing (Figure 14.B) [92]. The mRNA capping enzyme (mRNA CE) is recruited to the transcription complex, catalyzing its reaction through the binding of the phosphorylated Ser 5 in carboxyl-terminal domain (CTD) of RNA polymerase II (Figure 14.C) [93]. The capping modification is helpful in the recognition and attachment of mRNA to the ribosome as well as protection from exonucleases.



A. CTDSP1 & RNAP II B. PCF11 & RNAP II C. mRNA CE-α & RNAP II

A. CTDSP1									P	S	Y	S	P	T	S	P	S							
B. PCF11									P	S	Y	S	P	T	S	P	S							
C. mRNA CE-α								T	S	P	S	Y	S	P	T	S								
RNAP II /1-1749	1691	Y	S	P	T	S	P	S	Y	S	P	T	S	P	S	Y	S	P	T	S	P	Q	Y	S	1713



D. Superimposition of the 3 bound MoRFs

Figure 14. The MoRF mechanism plays a role in mediating interactions involving the CTD of RNA polymerase II. (A) CTD small phosphatase 1(with ι-MoRF), (B) protein 1 of cleavage and polyadenylation factor I (with ι-MoRF), (C) mRNA capping enzyme alpha subunit (with ι-MoRF), and (D) Similar bends near Pro 1700 occurs in all three bound MoRFs. In the sequence alignments, residues in red indicate residues with PTMs in PDB.

The three bound MoRFs in RNAP II all seem to exhibit a bend at a similar location. To gain greater insight, these three MoRFs were structurally aligned (Figure 14.D). Two of the MoRFs (bound to PCF11 and mRNA CE-α) show very similar backbone traces with bends at P1700. The third MoRF (bound to CTDSP1) also shows a bend near P1700, but the backbone trace and location of the bend relative to P1700 are different from the other two examples (Figure 14.D).

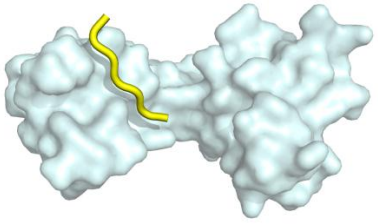
Since these sequences typically contain just one MoRF binding site for multiple partners, this raises the possibility that partner competition for the single site could be an important regulatory feature these binding interactions. In contrast, for the CTD of RNAP II, the MoRF sequence is repeated more than 50 times. These MoRFs may adapt different structures as they bind alternative partners. The interplay between partner competition and repeated binding sites may provide a mechanism for subtle and tunable regulation of MoRF / partner interactions.

The MoRF in Histone H3, which contains the maximal number of partners in our dataset, interacts with nine structurally different partners using residues from 2 to 22 in the sequence (Figure 15). Even though all nine MoRFs are classified as coiled structures, some residues within the MoRF region form helical or strand-like structures upon binding to the different partner proteins. Among the nine binding partners of the N-terminal tail of histone H3, there are several enzymes that are implicated in post-translational modifications. This N-terminal tail that protrudes from the globular nucleosome core can undergo several different types of epigenetic modifications that influence cellular processes. These modifications include the covalent attachment of methyl or acetyl groups to lysine and arginine amino acids and the phosphorylation of serine or threonine.

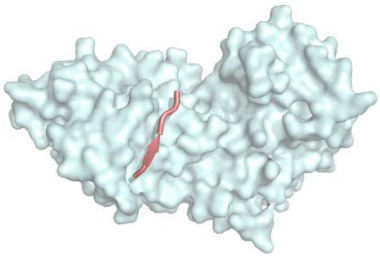
Some of these modifications are included in our data set and characterized in Figure 15 (with the modified residues marked in red).

The double Tudor domain of JMJD2A, a Jmjc domain-containing histone demethylase, binds methylated Lys 5 on Histone H3. This complex functions as a transcription repressor (Figure 15.A) [94]. The DNA-methyltransferase 3-like (DNMT3L) protein recognizes the histone H3 tails with unmethylated Lys 5 and stimulates de novo DNA methylation by engaging the DNMT3A2 molecule (Figure 15.B) [95]. The WD-repeat protein 5 (WDR5) is a core component of SET1-family complexes that achieve transcriptional activation via methylation of histone H3 on Lys 5 (Figure 15.C) [96]. The recombination activating gene (RAG) 2 contains a plant homeodomain (PHD) that recognizes histone H3 methylated at Lys 5 and influences V(D)J recombination (Figure 15.D) [97]. Histone demethylase LSD1 regulates transcription by demethylating Lys 5 of histone H3 (Figure 15.E) [98]. A substrate-like peptide was generated by a K5M mutation (marked in gray in Figure 15) because this mutation led to 30-fold increase in binding affinity thereby helping to stabilize the complex. Phosphorylation at Ser 11 of histone H3 enhances GCN5 histone acetyltransferase (HAT) mediated Lys 15 acetylation, promoting transcription (Figure 15.F) [99]. The 14-3-3 isoforms present a class of proteins that mediate the effect of Ser 11 phosphorylated histone H3 (Figure 15.G) [100]. The jumonji domain of JHDM3A (JMJD2A) catalyzes the demethylation of di- and tri-methylated Lys 10 and Lys 37 in histone H3 (Figure 15.H) [101]. DIM-5 is a histone H3 Lys 9 methyltransferase that is essential for DNA methylation (Figure 15.I) [102].

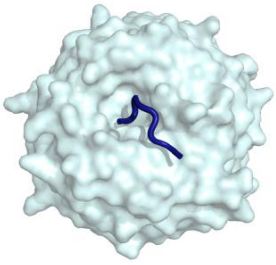
Figure 15.J summarizes the results of disorder/order predictions, potential interacting regions and annotated post-translational modification (PTM) sites in UniProt in human histone H3. In general, H3 has a central structural region (residue 58-132) that matches to a Pfam family (histone: core histone H2A/H2B/H3/H4) and a long N-term disordered tail (around 38-48 residues in length). A similar disorder/order estimate was given by PONDR VSL2B. Within current 294 PDB entries related to human histone H3 (27-Mar-12), 40 complexes were found to include H3 fragments (MoRFs) between residue 2 to 34. This N-terminal binding region was not recognized by both MoRF1 and MoRF2 predictors [45,103], but we claim the reasons may be because these two predictors were built specifically for helix MoRFs, not coil MoRFs like the ones in H3. Figure 15.A-I show the nine MoRFs found in the same region are all coil MoRFs. Part of the binding region can be predicted by ANCHOR [47] while the entire region can be found by MoRFpred [48] method. Based on the sequence annotations of UniProt database, most PTM sites of H3 are located in the N-terminus of H3, implying the functionally regulation sites may highly tie with MoRFs within disordered regions. Otherwise, it is very likely that there are 2 to 3 MoRFs tandem one by one in this case (e.g. 1-7; 8-14; 15-22).



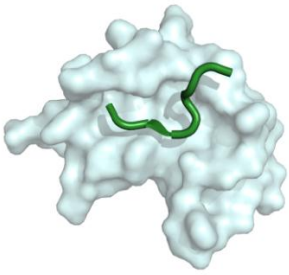
A. JMJD2A & H3



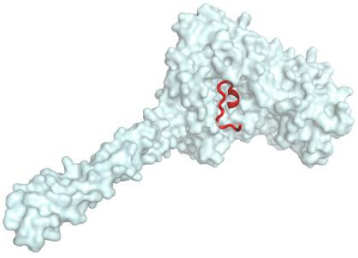
B. DNMT3L & H3



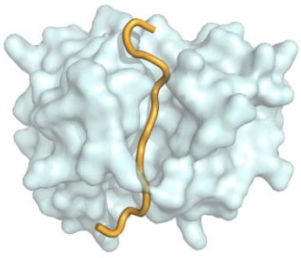
C. WDR5 & H3



D. RAG2 & H3

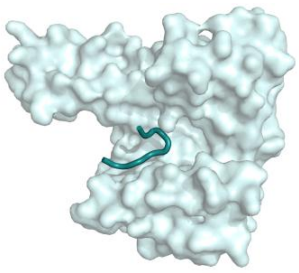


E. LSD1 & H3

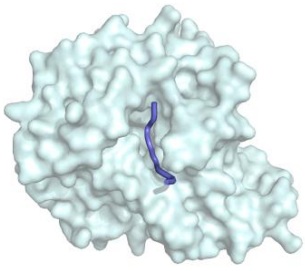


F. HAT A1 & H3

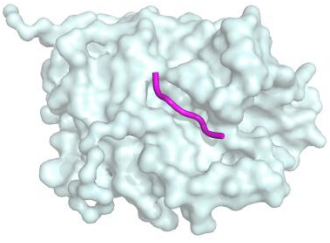
A. JMJD2A								A	R	T	K	Q	T	A											
B. DNMT3L								A	R	T	K	Q	T	A											
C. WDR5								A	R	T	K	Q	T	A	R										
D. RAG2								A	R	T	K	Q	T	A	R	K									
E. LSD1								A	R	T	M	Q	T	A	R	K	S	T	G	G	K	A	P		
Histone H3/1-136	1	M	A	R	T	K	Q	T	A	R	K	S	T	G	G	K	A	P	R	K	Q	L	A	T	K	A	25
F. HAT A1											A	R	K	S	T	G	G	K	A	P	R	K	Q	L	A	
G. 14-3-3											A	R	K	S	T	G	G	K							
H. JHDM3A									T	A	R	K	S	T	G	G									
I. DIM-5											A	R	K	S	T	G	G								



G. 14-3-3 & H3



H. JHDM3A & H3



I. DIM-5 & H3

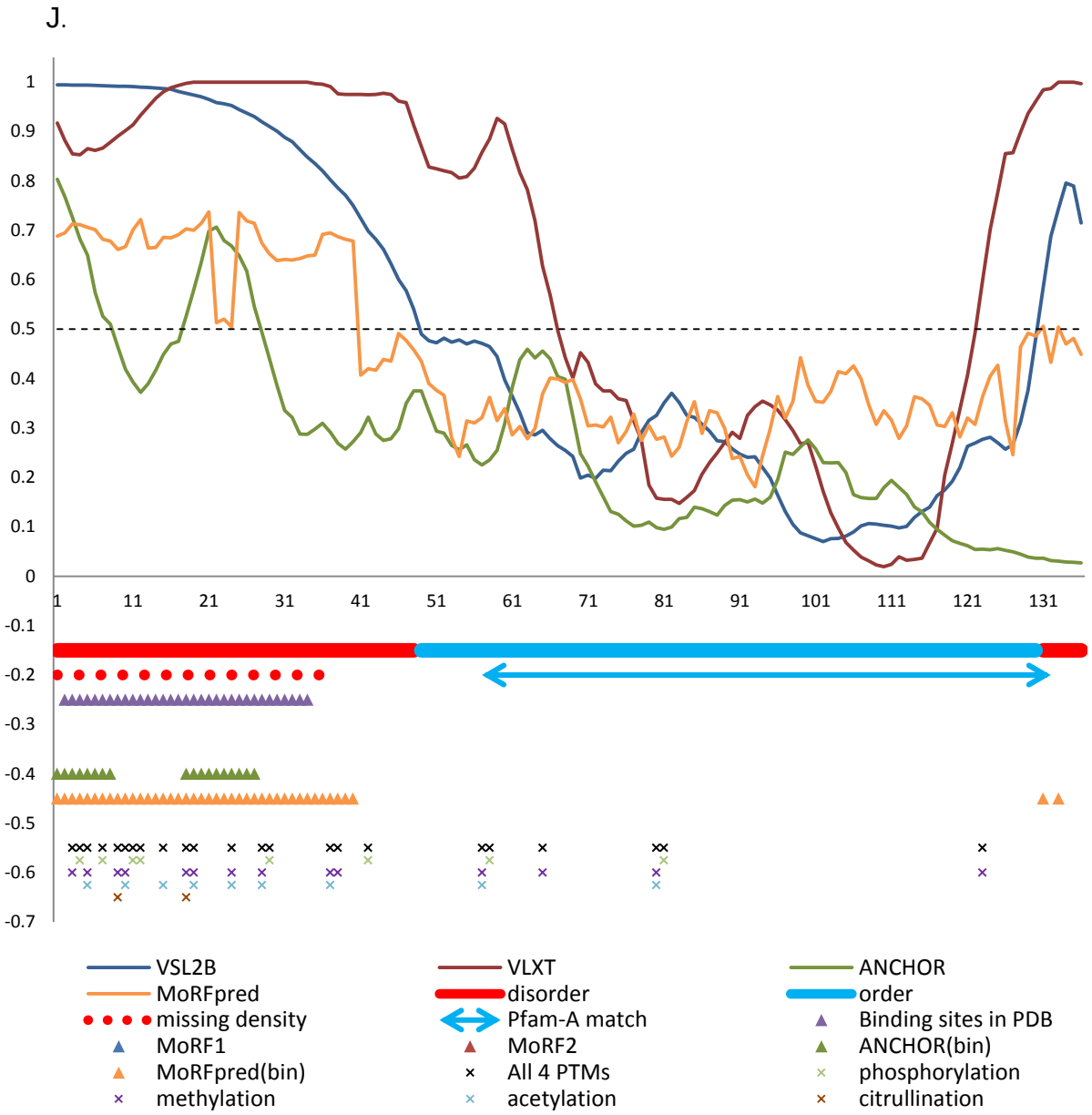


Figure 15. Nine different binding partners of ι-MoRFs in the N-terminus of histone H3. Its partners include (A) Jumonji domain-containing protein 2A, (B) DNA-methyltransferase 3-like, (C) WD-repeat protein 5, (D) VDJ recombination-activating protein 2, (E) lysine-specific demethylase 1, (F) Histone acetyltransferase (HAT A1), (G) 14-3-3 protein zeta/delta, (H) Jmjc domain-containing histone demethylation protein 3A, and (I) Histone H3 methyltransferase DIM-5. (J) Schematic diagram of histone H3 protein shows its predicted and validated disordered tails and a central folded domain. Structural data and various disordered binding site predictors reveal the potential binding regions of H3 are highly associated with posttranslationally modified sites. The residues in red in A-I are PTM sites in PDB, and the methionine in gray is residue that was mutated for the structural study. The annotated PTM sites on the entire H3 in J is from UniProt.

3.1.3. Alternative Splicing and Posttranslational Modifications in One-to-Many Binding

Our 23 MoRF examples of one-to-many binding comprise a special set, containing partners with little sequence similarity that bind to MoRFs with identical sequences. This approach is distinct from the concept of structural compensation or coadaptation, for which mutations on one partner are linked to compensating mutations on the partner [104]. It would certainly be possible to lift the requirement of MoRF sequence identity to thereby study coadaptation in complexes involving disordered proteins. Indeed, we have work in progress along these lines for a few specific examples to determine whether coadaptation between two structured proteins is different from coadaptation between structured proteins and MoRFs.

There have been several previous bioinformatics investigations of large numbers of IDP-involving protein-protein interactions at a high level, without paying attention to the structural details [45,103,105,106]. Instead, our approach here is to investigate fewer MoRF examples, but in greater in detail in order to develop a deeper understanding of how intrinsically disordered proteins can alter their conformations so as to be able to bind to structurally distinct partners. Our observations demonstrated that, in general, conformation flexibility allows for both subtle and complex structural variation, thereby enabling the same sequence to transform onto the diverse and distinctively shaped binding sites provided by their partners.

The MoRFs collected and grouped into one cluster herein are typically gathered from different organisms. As suggested by others, through parallel or convergent

evolution, such MoRFs can exist as conserved functional motifs or regions among various species, such as human, mouse, yeast, E. coli, or even viruses [107].

As pointed out previously [105], such short linear motifs are amenable to convergent evolution due to the limited number of mutations that are necessary for the generation of a useful motif. In fact, motifs are commonly used as adding new functional modules within a proteome, especially in higher eukaryotes [108]. These short functional linear motifs are hypothesized to have higher levels of conservation, to frequently evolve convergently, to preferentially occur in disordered regions and to often form a specific secondary structure when bound to interaction partners [107]. This observation fits in with the conception that alternative inclusion of exons in different tissues provides functional diversity of proteins. In fact, embedded conserved binding motifs and post translational modification sites are both rich in tissue-dependent protein segments [109]. The tissue-dependent spliced regions have higher percentage of protein disorder that likely form conserved interaction surface and participate significantly more protein interactions [110].

Among the 23 MoRFs in our dataset, three MoRFs (TRAP220, Bim and amyloid A4 protein) were annotated in UniProt to be located in alternatively spliced regions. Alternative splicing has the potential to add or delete an entire MoRF region. In addition, MoRF-related functions could be modulated by alternative splicing by changing the expression patterns, localization and regulation. These complex mechanisms could lead to broad functional and regulatory diversity. For example, pro-apoptosis protein Bim has 17 isoforms. Its predominant three isoforms, BimEL, BimL and BimS, all have the MoRF region (BH3 ligand) “DMRPEIWIAQELRRIGDEFNAYYAR”, which is

responsible for binding selectivity for their pro-survival protein binding targets and starting Bcl-2 regulated apoptosis. Those Bim isoforms lacking the BH3 ligand, e.g. Bimβ1-7, also lack pro-apoptotic activities.

Two additional MoRFs were reported to have alternative splicing events based on studies of the tissue-specific splicing exon data set [109]. A MoRF region from nuclear receptor corepressor 2 is specifically expressed in only 1 of 14 tissue types. As was pointed out [109], the tissue-specific alternative splicing that leads to presence and absence of binding sites in disordered protein regions leads to the "rewiring" of protein-protein interaction (PPI) networks, and may therefore contribute fundamentally to tissue development. It would be very interesting to develop models for the alterations in PPI networks in different tissues that arise from alternative splicing, but unfortunately the partners for the tissue-specific MoRFs are simply not known.

In a previous study, we found that alternatively spliced regions of RNA code for protein disorder much more often than for regions of structure, and we showed that such alternative splicing could lead to inclusion or exclusion of binding sites within the disordered regions [9]. Interestingly, of the human MoRFs studied here, 50% (4 of 8) are in exon regions that have been identified as included or excluded by alternative splicing. The discussion in the previous paragraph suggests that a concerted effort should be made to identify additional MoRFs that map to tissue-specific alternatively spliced regions and to identify their partners as well.

In our previous study of the carboxy terminal tail of p53 bound to 4 different partners, we noticed that two of the complexes were distinguished by having PTMs, namely lysine acetylations for both examples. Furthermore, the acetate groups both

became buried in the interfaces between the two MoRFs and their respective partners [46]. In this study we discovered that differences in PTMs occur commonly when MoRFs bind to alternative partners. Furthermore, this use of modified side chains to bind to one of two partners is most common when the two partners are structurally distinct. Indeed in this study, of 13 MoRFs containing PTMs, 11 involve MoRFs that bind to differently folded partners, thus providing additional observations in support of this concept. Finally, the chemical group added via the modification is typically found buried or partially buried in the interface between the MoRF and its partner, which strongly suggests that PTM provides an important part of the signal for the MoRF to bind to an alternative partner.

Phosphorylation occurs much more often in intrinsically disordered as compared to structured regions of proteins [111,112]. Recently, several other types of PTM have been shown to prefer disorder over structure [113]. The results presented herein suggest that such a modification can be used to change the partner preference of a given MoRF, thus leading to switching the connections of a protein-protein interaction network.

3.2. Many-to-One Binding

A total of 4368 binary protein complexes were collected from the Protein Data Bank for which two or more peptides of different amino acid sequences were bound to the same (100% sequence identical) globular protein partner, a type of interaction that we call many-to-one binding. These peptides, which are embedded within putative intrinsically disordered protein (IDP) regions and which we call molecular recognition features (MoRFs), were restricted to be of length 5 to 25 residues. Two distinct binding profiles were identified in the collected many-to-one binding dataset: independent and

overlapping (varying from similar to intersecting). Within a subset of 139 selected protein-protein interactions, 72 or 51.8% of MoRF binding sites are similar within their own cluster. For this similar binding profile, the distinct MoRFs interact with almost identical binding sites on the same partner. Next, 33 or 23.7% of the MoRF binding sites are independent. For this independent binding profile, the MoRFs within the same cluster interact with completely different parts of the same binding partner. Finally, 34 or 24.5% of the MoRF binding sites intersect without being highly similar. For this intersecting binding profile, the binding sites contain both common and unique interaction residues. Relatively higher sequence conservation is noted for those partner interfaces with similar binding residues. Further analysis of the sequence and structural changes within these three groups indicate how an IDP's flexibility allows different segments to adjust to similar, independent, and intersecting binding pockets.

Table 6 summarized all the procedures to collect our many-to-one binding dataset. First of all, 8084 short binding MoRFs having 5 to 25 residues were found in the PDB as of June 19, 2012. Within the initial MoRF set, 7064 of them have interactions likely to be of biological significance as estimated by the criterion of a buried surface area larger than 400 Å2. There are 6835 of these interaction complexes having folded binding partners whose sequence lengths are more than 40 residues. In order to identify peptides bound to the same partners, partner sequences were mapped back to their parent sequences. Here, 4612 partner sequences exhibited an exact match in Universal Protein Resource (UniProt) sequence database. 4368 partners were observed to overlap with at least one other protein, thereby leading to 514 distinct partner sets. After removing identical sequences with just one partner, 384 clusters with 2081 MoRFs were assembled. In other words,

each globular domain on average associates with 2081/384 = 5.4 different binding sequences.

Table 6. Description of many-to-one MoRF dataset

Data set	MoRFs/ Partners	Clusters	MoRFs per cluster
Initial MoRF dataset (5-25) [a]	8084		
MoRF dataset with biological interaction (>400Å^2) [b]	7064		
Partner dataset with sequence length (>40)[c]	6835		
Partner dataset mapped to UniProt sequence database	4612		
Partner dataset with overlapped region in mapping [d]	4368	514	8.50
Partner dataset without 100% sequence identity in MoRF	2081	384	5.42

[a]MoRFs with 5 to 25 residues are the focus of this study.
[b]400 Å^2 cutoff was set to filter out the spurious interactions caused by crystal contacts.
[c]Binding partners of MoRF are supposed to be globular proteins having more than 40 residues to fold into a certain conformation.
[d]Partners having one or more overlapping residues with each other.

3.2.1. Peptide-Protein Interactions and Protein-Protein Interactions

Rather than attempting a detailed analysis of 2081 complexes grouped into 384 clusters, we elected to study a smaller number in greater detail, thereby making one-by-one visual inspection of each complex much more practical. In the end, we turned our attention to 160 globular proteins bound to 909 MoRFs, giving an average of 909/160 = 5.7 sequences bound to each globular domain. This group was used to characterize binding profiles. Furthermore, 21 peptide-protein interactions were set apart as a special subset since these interactions occur only after the MoRFs are enzymatically chopped

from their parent proteins. These MoRF partners are all immune-related and the MoRFs themselves have less sequence conservation with each other. The 21 immune-related PPIs and the remaining 139 PPI clusters were collected and inspected leading to their inclusion in Table 7. The 384 clusters of Table 6, the 21 immune-related clusters, and the remaining 139 clusters are compared with respect to their partner-number distributions in Figures 16 and 17.

Table 7. Description of 2012 and 2008 MoRF datasets. 21 immune-related PPIs and 139 nonimmune-related PPIs.

	2012_dataset	2008_dataset	2008_immune	2008_nonimmune
100% Quartile (max)	85	48	48	20
75% Quartile	5	6	19	6
50% Quartile (median)	3	3	4	3
25% Quartile	2	2	2	2
0% Quartile (min)	2	2	2	2
average	5.42±7.86	5.68±6.79	13.05±14.64	4.57±3.55
sum	2081	909	274	635
count	384	160	21	139

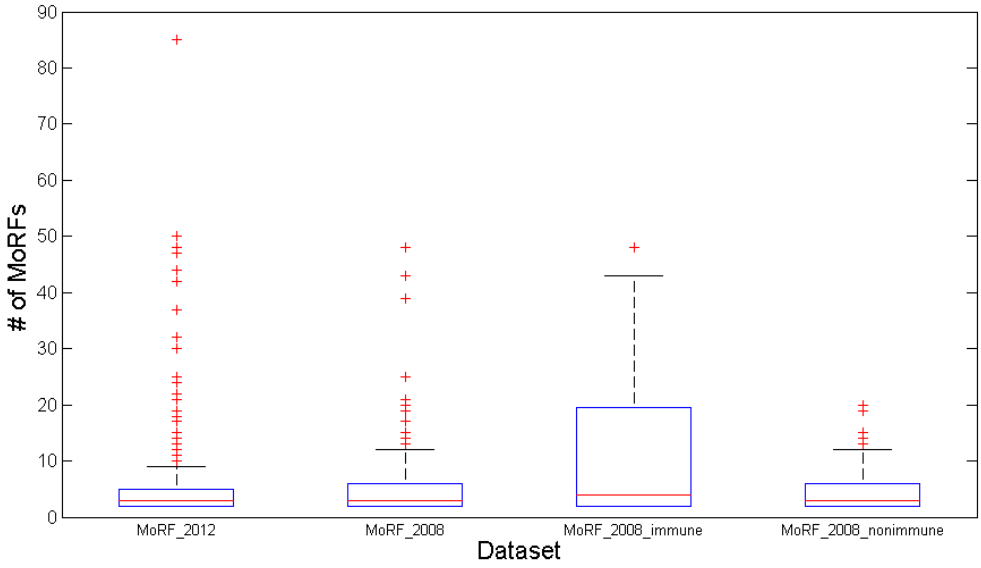


Figure 16. The boxplot of 4 different datasets: MoRF_2012, MoRF_2008, MoRF_2008_immune, MoRF_2008_nonimmune.

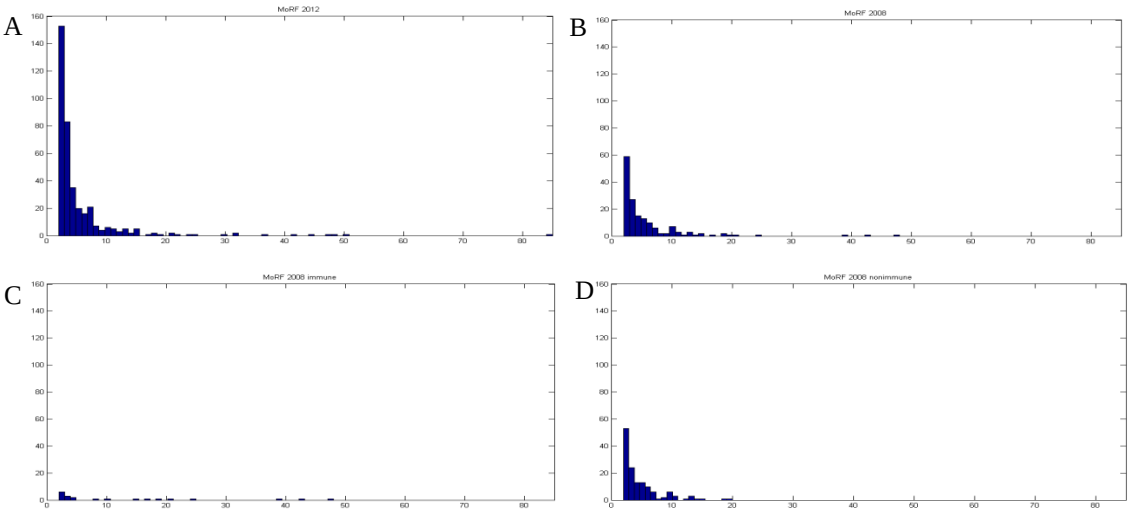


Figure 17. X-axis implies the counts of MoRFs within a cluster. Y-axis implies the counts of each cluster. The diagrams illustrated the data variation in different datasets as below:

(A) MoRF dataset collected in 2012
(B) MoRF dataset collected in 2008
(C) Peptide-Protein interactions in 2008 dataset
(D) Protein-Protein interactions in 2008 dataset

3.2.2. Binding Profiles: Independent and Overlapping (Similar vs. Intersecting)

Using one-by-one visual inspection of the 139 protein-protein interactions and 21 peptide-protein interactions, two main binding profiles were observed to describe many-to-one interactions (Figure 18). The MoRF binding pocket on the partner side may be located in separated parts of the same binding partner (independent binding pocket) or contain both common and distinct interacting residues (overlapping binding pocket). The volume overlap ratio can range from almost 0% (intersecting binding pocket) to 100% (similar binding pocket).

Table 8 gives a summary of each interaction category and the sequence identities of binding MoRFs. INDEL MoRFs are a set of similar MoRFs with only amino acid insertion or deletion on the terminal side within one cluster. On the other hand, mutative MoRFs have at least one amino acid difference as compared to the other MoRFs in the same cluster.

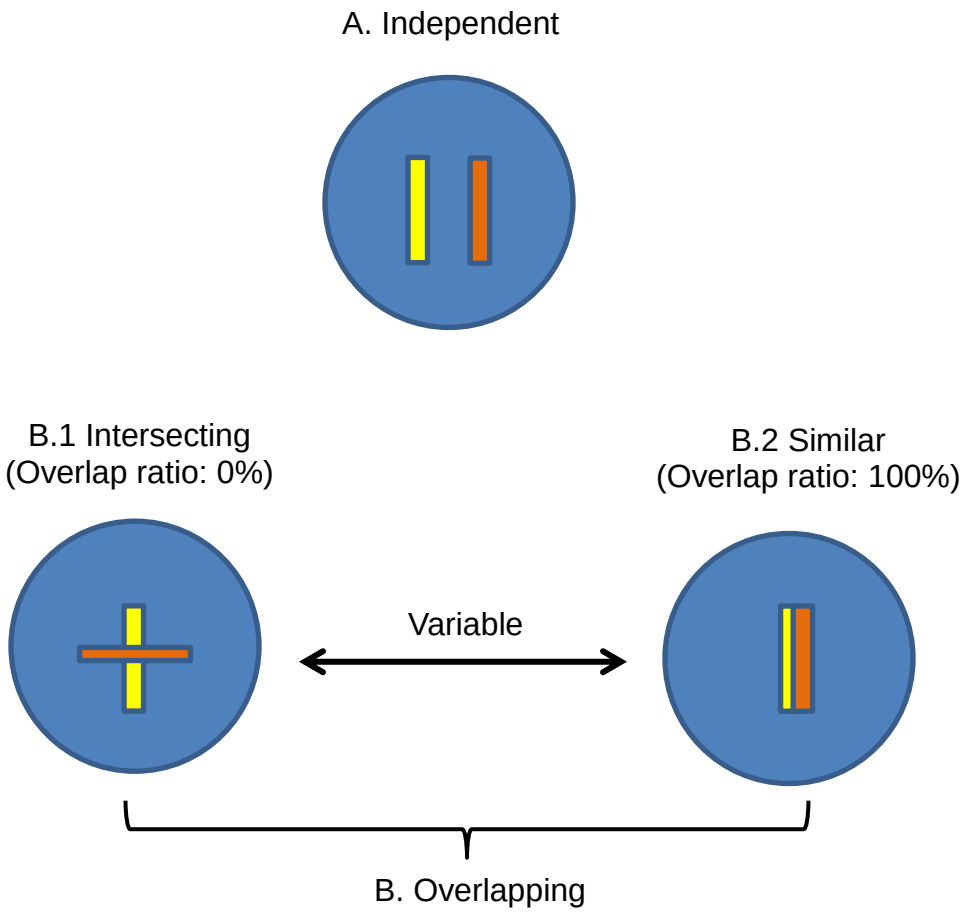


Figure 18. Many-to-one binding can be classified into two main categories: (A) independent and (B) overlapping binding pockets as the overlapping extent of a MoRF pair can range between (B.1) intersecting (overlap ratio: 0%) and (B.2) similar (overlap ratio: 100%).

Table 8. Peptide-protein interactions and protein-protein interactions in many-to-one MoRF dataset.

Interaction Type		Clusters	Identity	Overlap ratio
Peptide-protein interactions[a]		21		
G1	Similar binding pockets with INDEL[b] MoRFs	0		
G2	Similar binding pockets with mutative MoRFs	15	0.29	0.42-0.72 (0.50)
G3	Intersecting binding pockets	2	0.20	0.18-0.32 (0.25)
G4	Independent binding pockets	4	0.35	0.00-0.48 (0.16)
Protein-protein interactions		139		
G5	Similar binding pockets with INDEL[b] MoRFs	24	1.00	0.42-0.80 (0.62)
G6	Similar binding pockets with mutative MoRFs	48	0.60	0.40-0.88 (0.54)
G7	Intersecting binding pockets	34	0.52	0.14-0.38 (0.29)
G8	Independent binding pockets	33	0.52	0.00-0.46 (0.17)

[a]Peptide-protein interactions include MHC, T-cell receptor and antibody related complexes.
[b]INDEL MoRFs have either insertions or deletions with all the other MoRFs in the same cluster.

Sequence identities of interacting partner binding residues and RMSD of aligned multiple complexes can't make a good distinction for the three main binding profiles. In order to quantify the similarity between MoRF fragments attaching to the partners, molecule volume calculations of each pair of MoRF were estimated. By dividing the volume of combined MoRF pairs by the volume of separate MoRF pairs, we can get the volume ratio from 0.5 to 1. The volume ratio was normalized from 0 to 1 to measure the extent that MoRF pairs overlap, and we termed the value as overlap ratio. 1 indicates two MoRF fragments are fully overlapped and 0 indicates two MoRF fragments are 100% spatially separated. The cluster would be determined as an independent binding pocket directly if there is one MoRF pair's overlap ratio is 1 within the cluster. 0.4 was used as a cut off for overlap ratio to define similar binging pockets and intersecting binding pockets. In summary, we obtained 15, 2, 4 and 72, 34, 33 for similar, intersecting and independent binding pockets in peptide-protein interactions and protein-protein interactions, respectively. The overlap ratio for independent binding has a wider range because there might be similar or intersecting binding within the same cluster. The TAZ1 domain was not included in our many-to-one binding dataset since its binding peptides, HIF1α and CITED2, are both exceeding the upper bound length criteria of MoRFs.

Figure 19.A-J illustrates the data distribution of the three main binding profiles. The X-axis is the overlap ratio and Y-axis is the MoRF counts. The results imply the binding profiles of those peptide-protein interactions (immune-related) are very different from the regular protein-protein interactions (nonimmune-related). Unfortunately, the current way we used to classify the three binding profiles still need to be improved a lot. We are looking for a better way to separate the three classes by other algorithms.

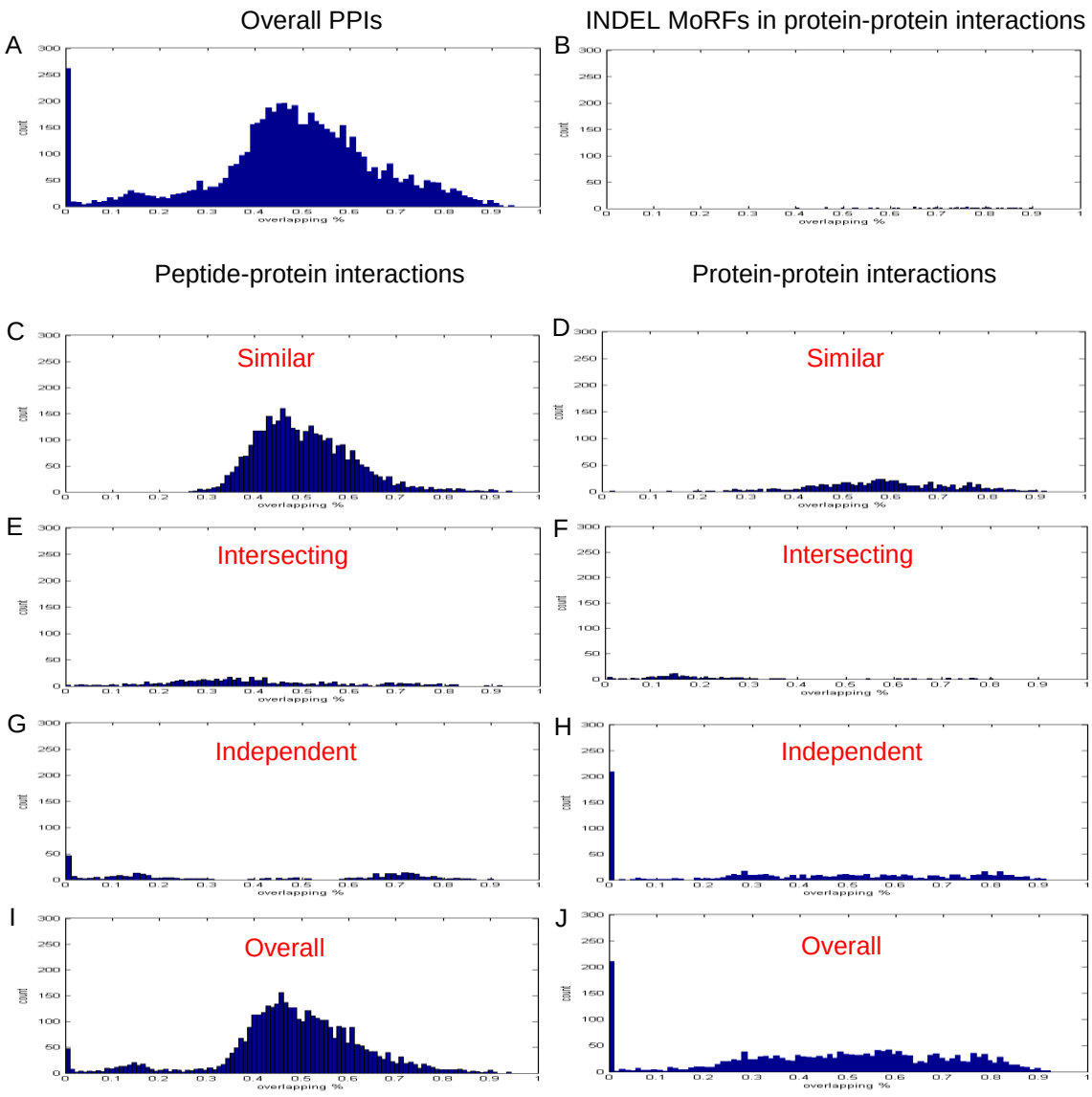


Figure 19. Histograms of pairwise overlap ratio based on all atoms calculations in the following datasets:

(A) 160 many-to-one binding examples in 2008 dataset
(B) Similar binding profiles with INDEL MoRFs in protein-protein interactions
(C) Similar binding profiles in peptide-protein interactions
(D) Similar binding profiles in protein-protein interactions
(E) Intersecting binding profiles in peptide-protein interactions
(F) Intersecting binding profiles in protein-protein interactions
(G) Independent binding profiles in peptide-protein interactions
(H) Independent binding profiles in protein-protein interactions
(I) All peptide-protein interactions (immune-related)
(J) All protein-protein intersections

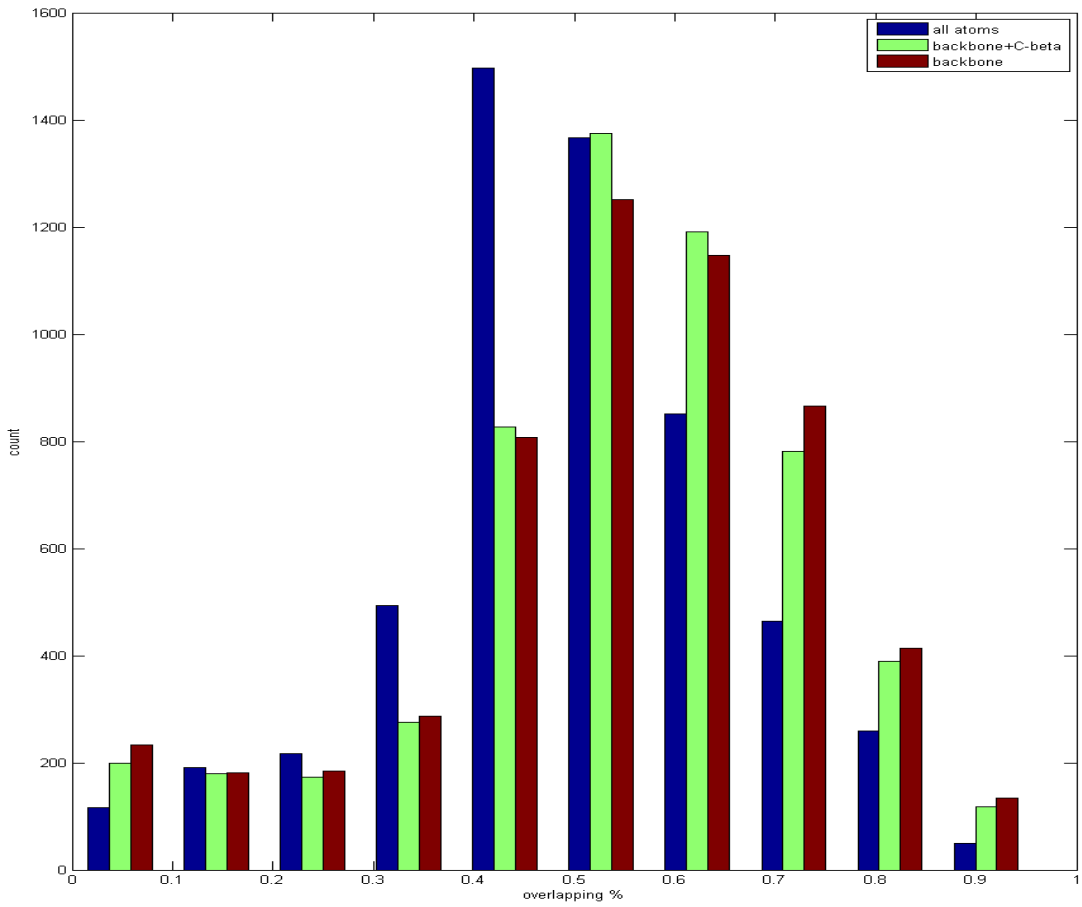


Figure 20. Histograms of pairwise overlap ratios based on all atoms (blue), backbone+C-beta (green) or backbone (red) calculations.

The overlap ratios were calculated in three different ways using various set of atoms, including all atoms (blue), backbone + C-beta (green) or only backbone (red). Two-sample Kolmogorov-Smirnov tests were performed to examine whether the overlap ratio values calculated by different approaches have the same distribution at 5% significance level. The results show side chain conformations have a significant effect on the overlap ratio calculations (Figure 20).

The null hypothesis that each pair of distributions is the same could be rejected with the following p-values for each comparison as shown below:

1. “all atoms” vs. “backbone + C-beta” => p-value: 1.4187e-68
2. “backbone + C-beta” vs. “backbone “ => p-value: 0.0704
3. “all atoms” vs. “backbone” => p-value: 6.2259e-72

3.2.3. Structurally Conserved MoRFs with Diverse Sequences

The conformations of the binding MoRFs with the identical partner were examined by comparing their secondary structure assignments (Table 9). Although MoRF sequences in each set are different, their folded secondary structures are highly conserved. 71% sets have identical secondary structure assignment, indicating dissimilar MoRFs tend to pack into a specific conformation upon binding to one partner. Coils are the majority of secondary structures on MoRFs (100 examples). Otherwise, 9 examples contain 3 secondary structure types and 38 examples have 2 secondary structure types, implying structurally diverse MoRFs also exist in part of our sets. In the peptide-protein interaction subset, the conformations of the MoRFs tend to converge toward coils. Within general protein-protein interactions, MoRFs bound to intersecting binding sites have more diverse structures than those binding to the similar binding sites. The INDEL MoRFs have more similar conformations than the mutative MoRFs. MoRFs with independent binding sites have the widest range of secondary structure types.

In addition, we performed a simple amino acid composition profile comparison by a web-based tool called composition profiler [114]. Figure 21 shows the comparison between our query datasets (e.g. MoRFs, iMoRFs) and reference datasets (e.g. PDB, DisProt).

Table 9. The combination of secondary structure types in the 160 partner sets.

MoRF Secondary structure	Clusters	G2	G3	G4	G5	G6	G7	G8
α + β + ι + Complex	0	0	0	0	0	0	0	0
α + β + ι	1	0	0	0	0	0	1	0
α + β + Complex	1	0	0	0	0	0	0	1
α + ι + Complex	4	0	0	0	0	0	1	3
β + ι + Complex	3	0	0	0	0	2	0	1
α + β	1	0	0	0	0	0	0	1
α + ι	15	0	0	0	0	7	4	4
α + Complex	2	0	0	0	2	0	0	0
β + ι	8	1	0	0	0	2	3	2
β + Complex	0	0	0	0	0	0	0	0
ι + Complex	12	0	0	0	3	4	1	4
α	10	0	0	0	4	4	0	2
β	3	0	0	0	2	1	0	0
ι	100	14	2	4	13	28	24	15
Complex	0	0	0	0	0	0	0	0

G1 to G8 imply distinct groups in two different types of interactions in Table 8.

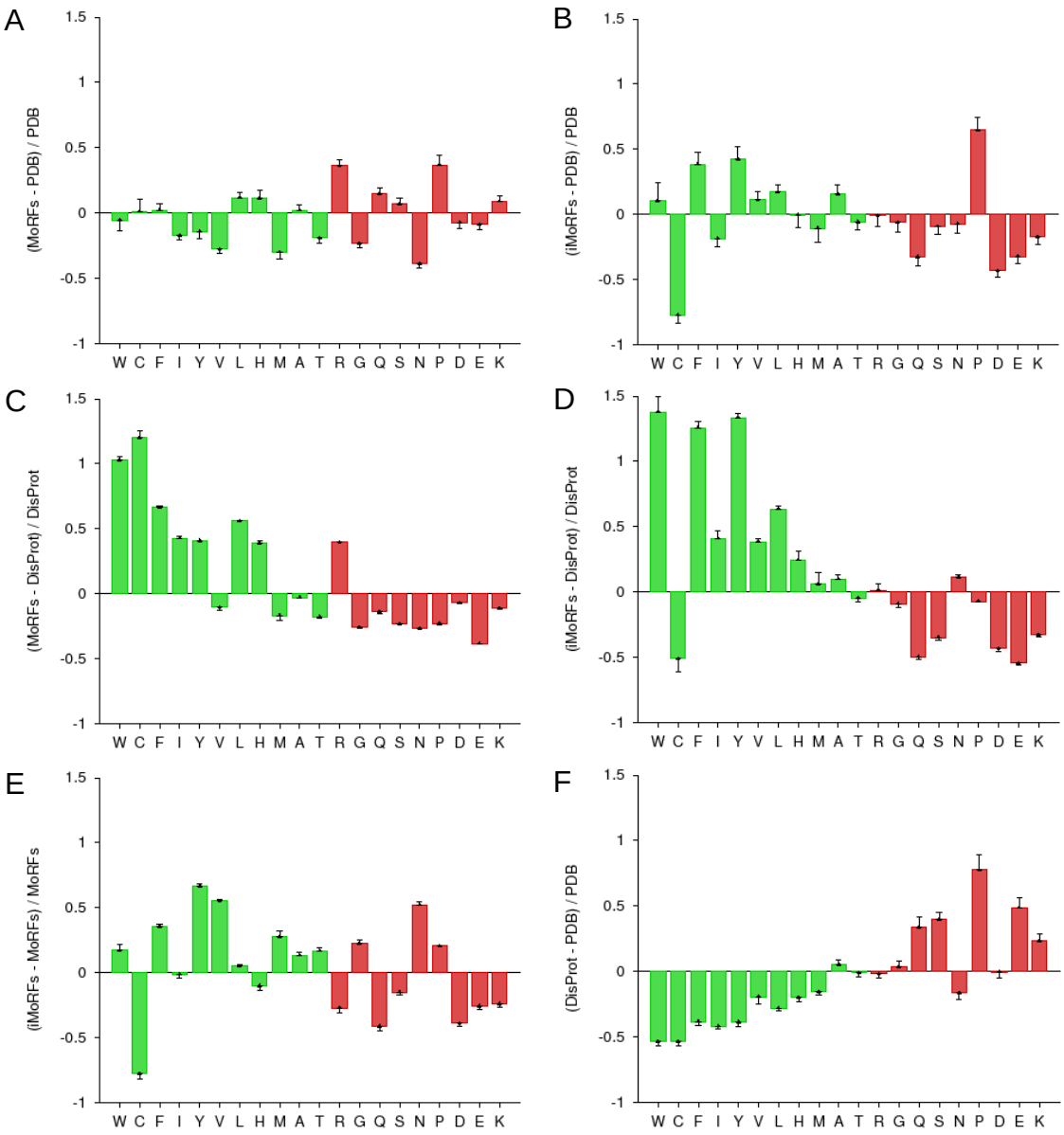


Figure 21. Amino acid composition comparisons between various datasets: (A) MoRFs – PDB, (B) iMoRFs – PDB, (C) MoRFs – DisProt, (D) iMoRFs – DisProt, (E) iMoRFs – MoRFs and (F) DisProt – PDB. iMoRF means immune-related MoRFs. Data and diagrams were generated by composition profiler (http://www.cprofiler.org/) [114].

3.2.4. Selected Many-to-One Case Studies

Several representative complexes with different binding profiles in peptide-protein interactions and protein-protein interactions were selected to delineate the many-to-one binding dataset.

(1) Peptide-protein interaction (immune-related) (Figure 22)

Figure 22.A depicts HLA class I histocompatibility antigen as a diverse MoRF set (with only 16% pairwise sequence identity between 39 MoRFs) but bind to a similar binding region. MHC molecules display a molecular segment called epitope for antigen presentation. IgG2A FAB fragment has an intersecting binding region interacting with 17 MoRFs sharing 19% sequence identity (Figure 22.B). Germline antibody FAB heavy chain binds to 2 MoRFs sharing 12% sequence identity without any overlapping their binding region (Figure 22.C).

(2) Protein-protein interactions (nonimmune-related) (Figure 23)

Six different binding profiles of protein-protein interactions were shown in Figure 23. Figure 23.A represents five INDEL MoRFs binding to a similarly overlapped binding site on WD repeat protein 5. 11 MoRFs with 42% sequence identity bind to a highly overlapped region of TNF receptor associated factor 3 in Figure 23.B. Ten distinct MoRFs (26% sequence identity) bind to an intersecting pocket on proteinase K having an intersecting binding pocket in Figure 23.C. An intersecting binding also appears in alpha-bungarotoxin bound to 10 MoRFs with 56% sequence identity. Two different patterns of MoRFs bind to two regions of chymotrypsin independently. Two patterns of MoRFs bind to two parts of focal adhesion kinase 1 separately.

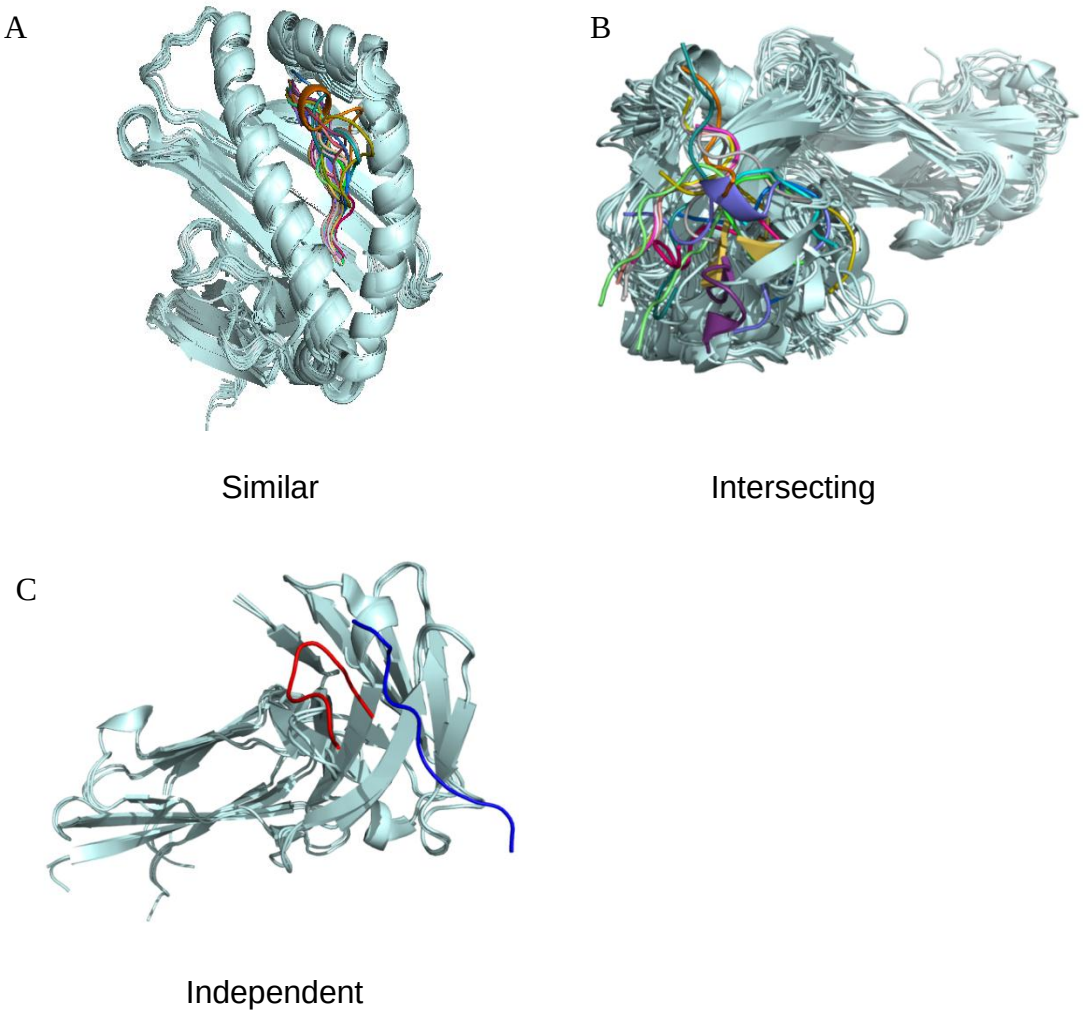


Figure 22. Peptide-protein interactions with similar binding pockets in (A) HLA class I histocompatibility antigen, intersecting binding pockets in (B) IgG2A FAB fragment and independent binding pockets in (C) Germline antibody 36-65 FAB heavy chain.

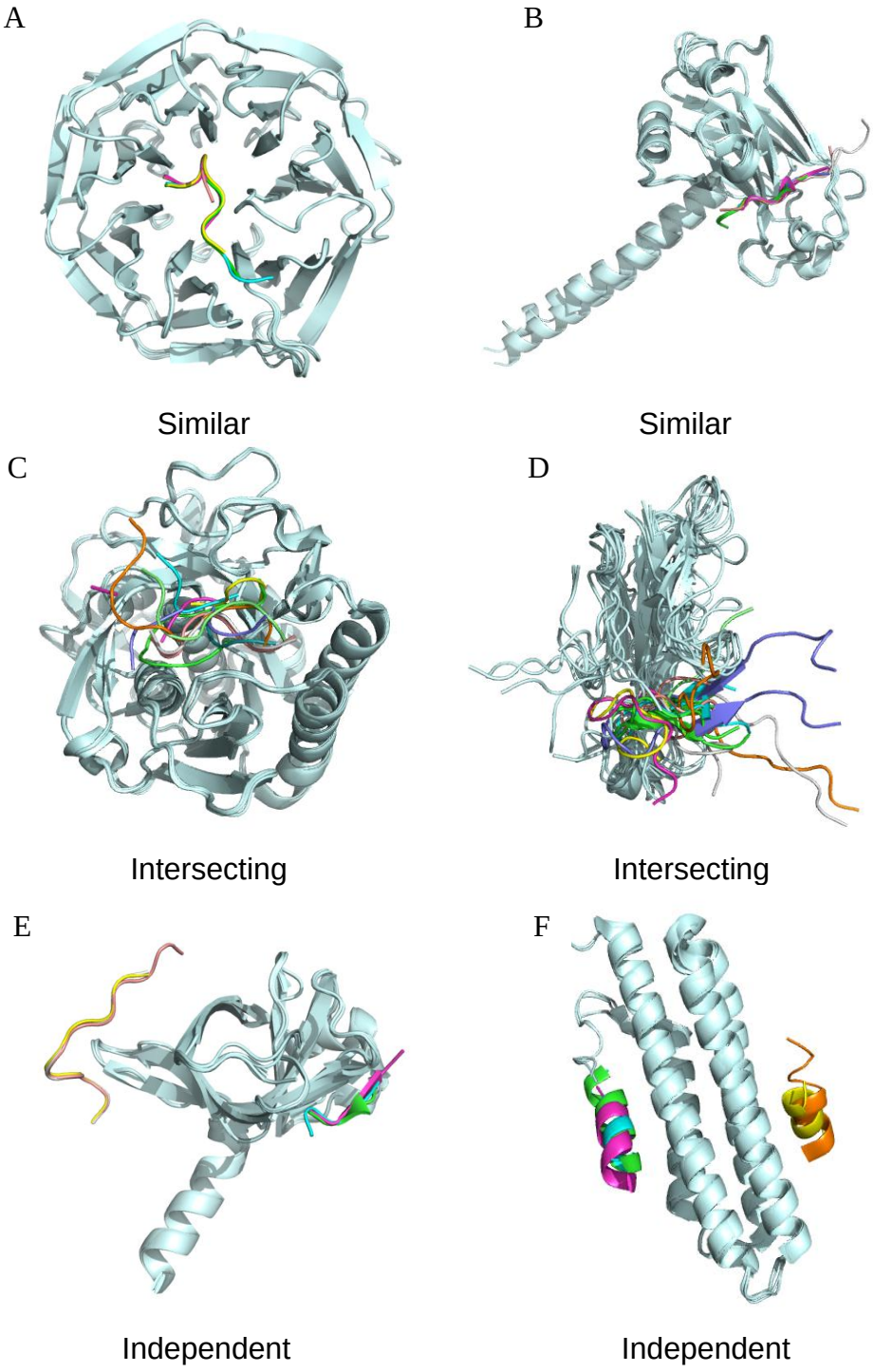


Figure 23. Protein-protein interactions with similar binding sites in (A) WD repeat protein 5 (B) TNF receptor associated factor 3, intersecting binding sites in **(**C) Proteinase K (D) Alpha-bungarotoxin and independent binding sites in (E) Chymotrypsin (F) Focal adhesion kinase 1.

3.2.5. Examples of Retro-MoRF and PP1-like MoRF

Retro-MoRF is a putative concept that a MoRF partner would also bind to the reversed sequence of an identified MoRF. The rationale behind the concept is that flexibility of protein disorder is sufficient to accommodate the altered geometry of a reverse sequence [115]. Since our many-to-one dataset collecting multiple disordered sequences binding to one common partner, we developed an algorithm and tried to search all the retro-MoRF cases in each cluster computationally.

Three poly-proline protein fragments were found to bind to the same pocket of actin regulatory protein profilin with both N-to-C terminal and C-to-N terminal directions (Figure 24). Profilin was demonstrated to bind to proline-rich peptides in two distinct backbone orientations like SH3 domain by X-ray crystallography data [116]. A poly-proline protein fragments usually fold into polyproline helix: either poly-pro I (PPI) or poly-pro II (PPII).

PPI helix is a right-handed helix with a much denser packing due to the cis isomers of its peptide bonds. PPI is rarer than PPII conformation because the cis isomer is not energy preferred. The dihedral angles for both conformations are close but not identical. There is no hydrogen bond within PPI or PPII helix. PPII helix is a left-handed helix with three residues per turn. PPII helices can be bound to SH3 domains. In addition to PPI and PPII, there are other proline-rich binding domains like WW and EVH1.

Type 1 protein phosphatase (PP1) is regulated by inhibitor-2 (I2) by forming a complex with three MoRF regions within disordered I2: residues 12-17, residues 44-56 and residues 130-169. Another two examples were found in our MoRF dataset which use the similar mechanism. FXDXF motif (residues 1460-1471) and WVXF motif (residues

1479-1490) from synaptojanin 1 bind to an appendage domain from adapter-related protein complex 2. Insulin B chain (residues 25-43) and insulin A chain (residues 90-109) bind to an insulin degrading enzyme (Figure 25).

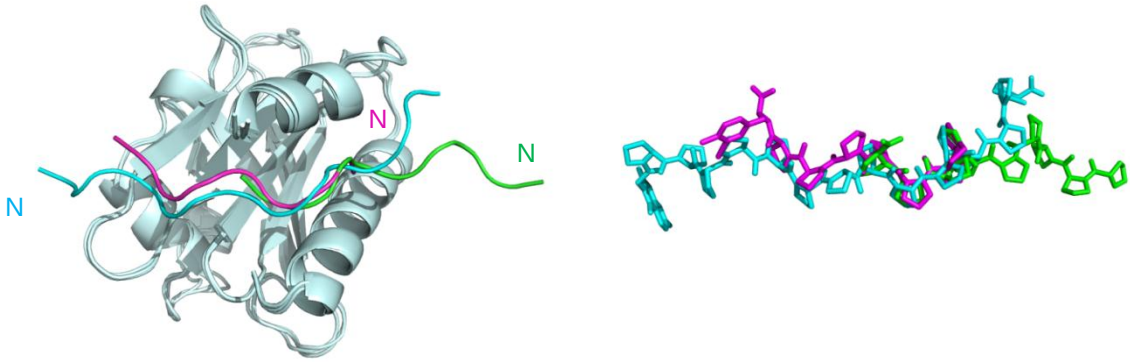


Figure 24. Poly-Proline containing Retro-MoRFs interact with the same binding pocket of Profilin.

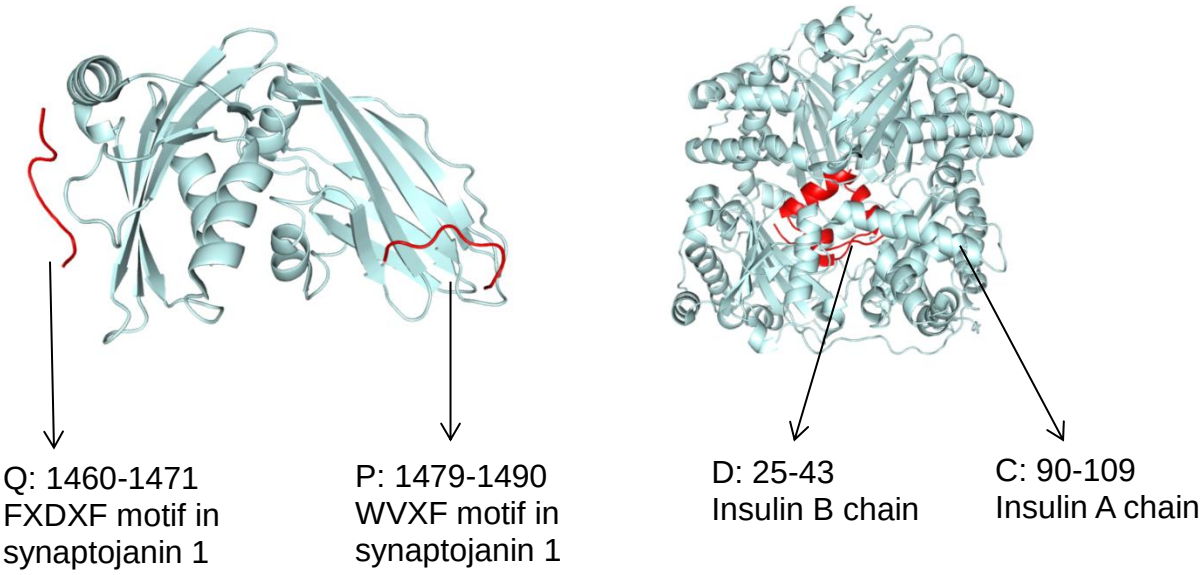


Figure 25. Different MoRFs from the same parent protein bind to a common partner.

3.3. Many-to-Many Binding

Besides one-to-many and many-to-one binding, we also found 12 interesting examples coexisting in both mechanisms, termed as many-to-many binding here. In each example, a common MoRF region interacts with different structured partners (one-to-many) which associate with diverse MoRF sequences simultaneously (many-to-one). Notice that all the structured partners of these 12 examples are belong to similarly folded partner category. Table 10 is a summarized list of the 12 many-to-many binding set we found.

Table 10. A summarized list of the 12 many-to-many binding examples.

	MoRFs	Num of SPs	SPs	Num of MoRFs
1	NCOA1	2	Estrogen receptor beta	2
			Peroxisome proliferator-activated receptor gamma	6
2	BAK peptide	2	BCL-XL	2
			M11L protein	0
3	NR 0B2-N-term	2	NR5A2	3
			Ancestral corticiod receptor	0
4	NCOA1 & 2	5	Androgen receptor	13
			Estrogen receptor	0
			NR1I3	2
			NR1I2	0
			Bile acid receptor	0
5	NR 0B2-C-term	2	Peroxisome proliferator-activated receptor gamma	6
			Androgen receptor	13
6	GRIM	2	Apoptosis 1 inhibitor	4
			Apoptosis 1 inhibitor	4
7	TRAP220	3	Vitamin D3 receptor	0
			Retinoic acid receptor, beta	0
			Retinoic acid receptor RXR-alpha	10
8	Latent membrane protein 1	2	TNF receptor associated factor 3	6
			Tumor necrosis factor receptor associated protein 2	11
9	NCOR2	3	Peroxisome proliferator activated receptor	4
			Estrogen-related receptor gamma	6
			Progesterone receptor	2
10	Amyloid beta A4	2	Disabled homolog 1	2
			X11	0
11	DNA repair protein RAD9	2	Protein kinase SPK1	5
			Probable regulatory protein embR	2
12	Nicotinic receptor	2	Alpha-bungarotoxin	10
			Long neurotoxin	0

CHAPTER 4

SCOP Folds of MoRF partners

4.1. Partner Folds Selection in Each MoRF Types

The binding diversity of protein disorder is a stumbling stone for modeling these cryptic disorder-order interfaces [117]. A typical example of MoRF in c-terminus of p53 can transform into helix, sheet or coil structures to adapt four different binding surfaces of partners with different folds. Strikingly, preliminary data from Dr. Sarah Bondos's laboratory reveals that Ultrabithorax (Ubx) protein prefers to interact with specific folds (Ref coming soon, PLOS ONE). Ubx is a Hox transcription factor and known to have intrinsically disordered regions regulating multiple gene regulations. Ubx coordinate multiple cell functions by associating with dozens of molecules. Evidence has shown that 22 of 29 Ubx binding partners can be assigned to only seven SCOP (Structural Classification of Proteins) folds, implying a possible hypothesis: partner topology may predetermine the molecular recognition between various partner bindings. These findings motivate us to computationally explore a large number of disorder-order interfaces by looking at the relationship between MoRF secondary structure type and structural topology of its potential binding partners.

Table 11 shows the SCOP fold classification of binding partners in each MoRF type (helix, sheet, coil and complex). There are totally 3148 MoRFs with 3750 SPs in our 2008 MoRF dataset. However, some partners' SCOP classifications haven't been assigned by SCOP database. Also, one partner may have multiple SCOP classifications

since there may be several different domains within one partner. Both issues increase the complexity of our hypothesis validation process.

Table 12 lists the major SCOP folds of MoRF binding partners that have greater than 2% proportion in each MoRF type. A trend was shown here that β-MoRFs prefer binding to all beta proteins (>52.4%) and α-MoRFs prefer to associate with all alpha proteins (44.2%).

Table 11. The SCOP fold classification of binding partners in each MoRF type.

	SPs	Partners in SCOP		Folds in SCOP	Non-redundant Folds in SCOP	Ratio
		No	Yes			
Helix	470	197	273	319	58	319/58=5.5
Sheet	139	47	92	105	23	105/23=4.6
Coil	2992	1026	1966	2999	145	2999/145=20.7
Complex	146	79	67	86	28	86/28=3.1
N/A in DSSP	3	0	3	3	1	3/1=3
Total	3750	1349	2401	3512	172	3512/172=20.4

Table 12. lists the major SCOP folds of MoRF binding partners that have greater than 2% proportion in each MoRF type.

	α-MoRF	P (%)	β-MoRF	P (%)	ι-MoRF	P (%)	Complex-MoRF	P (%)
Major SCOP Folds of SPs	a.123	32.6	b.47	14.3	b.1	27.8	b.1	11.6
	b.1	9.1	b.1	11.4	d.19	14.3	a.123	10.5
	a.39	8.5	b.2	10.5	b.50	4.6	b.47	9.3
	c.23	4.4	b.50	9.5	a.118	3.4	b.50	8.1
	c.37	3.4	a.39	7.6	d.93	2.8	b.36	7.0
	a.24	3.1	b.36	6.7	d.144	2.6	b.22	5.8
	b.121	3.1	e.1	6.7	a.74	2.1	b.55	5.8
	d.5	3.1	d.131	5.7	b.47	2.0	b.61	5.8
			d.5	5.7			a.11	4.7
			d.39	3.8			d.15	4.7
			d.144	2.9			g.7	3.5
							a.39	2.3
							c.55	2.3
							g.52	2.3

Table 13. Classification of MoRF binding partners based on 4 different levels in SCOP.

	Class	Fold	Superfamily	Family
Helix	9	58	64	73
Sheet	5	23	24	25
Coil	9	145	179	222
Complex	6	28	29	34

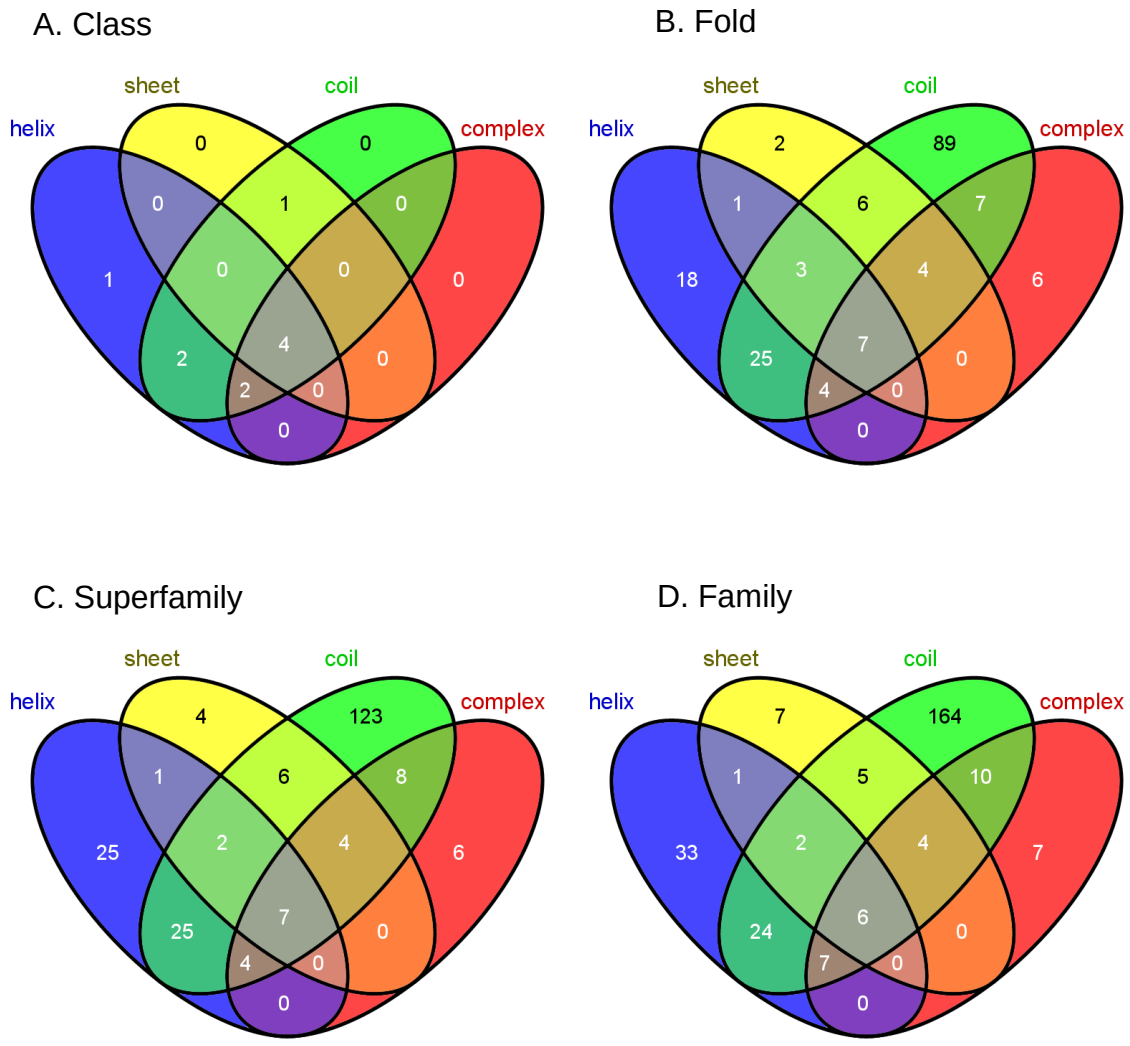


Figure 26. Venn diagrams of SCOP classification of MoRF partners in (A) class, (B) fold, (C) superfamily and (D) family levels for each MoRF type.

Table 13 summarizes the statistics of SCOP classification of MoRF binding partners in class, fold, superfamily and family levels. The same results were visualized and represented in another way showing in the Venn diagrams in Figure 26. Some more detailed information is listed in Table 14 (in SCOP class and fold level).

Table 14. Statistics of SCOP classification in (A) class and (B) fold levels of MoRF binding partners for each MoRF subgroup.

	A. Class		B. Fold	
Subgroup	Count	Member	Count	Member
h	1	k	18	a.1,a.12,a.2,a.7,a.70,a.91,b.84,b.93,c.120,c.49,d.104,d.129,f.1,f.23,f.32,h.4,i.22,k.21
s	0		2	e.1,g.54
c	0		89	a.102,a.129,a.133,a.144,a.146,a.158,a.20,a.202,a.22,a.23,a.246,a.29,a.35,a.48,a.58,a.60,a.74,a.8,a.98,b.103,b.119,b.130,b.136,b.18,b.26,b.3,b.57,b.62,b.68,b.70,b.72,b.82,b.85,b.86,b.9,c.1,c.104,c.111,c.14,c.15,c.17,c.2,c.34,c.45,c.47,c.52,c.56,c.57,c.66,c.69,c.7,c.72,c.8,c.93,c.94,d.108,d.126,d.135,d.136,d.142,d.159,d.165,d.166,d.169,d.17,d.195,d.198,d.20,d.200,d.223,d.33,d.54,d.56,d.58,d.88,d.9,d.95,e.28,e.3,e.45,e.8,f.4,g.16,g.17,g.27,g.3,g.44,h.1,i.6
com	0		6	a.223,b.108,b.81,b.91,d.231,g.1
h+s	0		1	d.5
h+c	2	f,i	25	a.118,a.24,a.28,a.4,a.59,a.71,b.121,b.29,b.34,b.40,b.42,b.69,c.23,c.31,c.51,c.62,d.105,d.109,d.110,d.185,d.42,d.86,d.92,g.14,g.41
h+com	0		0	
s+c	1	e	6	b.4,b.76,d.19,d.3,d.39,d.93
s+com	0		0	
c+com	0		7	b.22,b.6,b.8,c.41,c.55,d.26,h.3
h+s+c	0		3	a.42,b.2,d.131
h+s+com	0		0	
h+c+com	2	c,h	4	a.123,a.66,b.61,c.37
s+c+com	0		4	b.36,b.50,g.52,g.7
h+s+c+com	4	a,b,d,g	7	a.11,a.39,b.1,b.47,b.55,d.144,d.15
	10		172	

CHAPTER 5

Conclusion

When the idea of hub-based "scale free" protein-protein interaction networks was first proposed [10], a News and Views article pointed out that such multiple interactions were unlike what had been studied up to that time and that an understanding of these multiple interactions would likely require the discovery of new concepts [12]. Our laboratory immediately tried to suggest that the new principle was likely the use of disordered proteins by means of coupled binding and folding, which had been previously suggested for protein-DNA interactions [118] as well as for one protein binding to several partners [104], but publication of these ideas for hub proteins was delayed somewhat [51]. By now there is strong evidence that, at least for many if not all hub-partner interactions, disorder plays an important role in enabling one protein to bind to multiple partners [17-19,51].

As we have shown here, one MoRF can bind to multiple partners and, through these sets of interactions, a small region of one protein can play a role in multiple signaling events, thereby affecting several different cellular functions. Limitations on the availability of multiple interactions in PDB led to a small dataset, which, nevertheless, showed consistent results in terms of residue conservation in the binding partners. By restricting sequence identity of binding partners to 25% we were hoping to find MoRFs binding to structurally diverse partners, but it turned out that in most cases the partner's folds were very much conserved, despite the sequence differences. Aromatic residue side chain re-orientation was shown to contribute significantly on binding interfaces in two

and three interactions when the backbone conformation remained constant. Overall, these results paint a more detailed picture of multiple interactions than what has been available to date, and support the notion of intrinsic disorder or structural adaptability as an important factor in the development of non-random protein interaction networks characterized by multiple binding partners.

Despite the biological and structure-function importance of these disorder-based multiple protein interactions, there have been surprisingly few studies indicating in detail how such multiple interactions are brought about. Investigating more examples in detail is needed to provide a clearer and more general picture of how the one MoRF sequence can bind to two or more different partners.

To broaden our understanding, it makes sense to study different collections of single proteins binding to multiple partners and to study multiple proteins binding to the same partner. Previously we investigated one segment binding to completely unrelated partners (e.g., one-to-many signaling) and a collection of unrelated disordered partners binding to a single binding site on one structured protein (e.g. many-to-one signaling) [46]. Rather than focusing on individual examples as before [46], here we studied a collection of MoRFs involved in binding to more than one partner.

A distinctive feature of this study is the partners to a given MoRF had a sequence identity less than 25% yet displayed the same overall fold. In general, sequence variations in one partner are frequently linked to sequence variations on the other partner, indicating structural compensation or coadaptation across the binding interface [104]. However, the interacting protein pairs we collected here are a special set in that only the partners show

amino acid substitutions in their sequences, whereas the MoRFs' sequences are unchanged.

The indispensability of hub proteins is apparent, as they appear to evolve more slowly and are more likely to be vital for survival [10]. MoRF-protein interactions likely have the combination of high specificity coupled with low affinity [45]. The latter property may facilitate the discovery of small drug molecules that block the interactions. This study and others like it have the potential for a new strategy for drug discovery, namely to search for molecules that selectively block certain protein-proteins interactions involving a given protein but not others, by taking advantage of different conformations in the different interactions. This would allow the development of drugs that target specific pathways or even particular pathways in particular tissue types.

Observations have been made that the residues in enzyme active sites tend to evolve more slowly than other parts of the same proteins [80]. We wondered whether the same trend would also be found in the binding sites of the structured partners. By analyzing the 11 sets of interactions that we collected, we found that, like the active-site residues of enzymes, the binding residues of the structured partners exhibited a higher conservation as compared to the non-binding residues.

A recent study showed that protein-protein interactions in which a disordered region binds to a structured partner often involves interactions between two aromatic groups, for which the aromatic residues are frequently not stacked but rather oriented in such a way that a hydrogen of one aromatic ring points towards the centers of the conjugated electron rings of the other [119]. In agreement with this study, many of the protein-protein interactions investigated herein do indeed involve interacting aromatic

residues, but specific examples without such aromatic-aromatic interactions were also found.

Interactions between globular proteins and MoRFs often contain disordered residues as part of the MoRFs [38]. Others have shown that, even though such local regions remain unstructured, they can still affect the overall binding affinity. Such "fuzzy complexes" thus bind to their partners without undergoing complete conversion to structure with the regions that remain disordered still contributing to the energetics of the interaction [120]. From what we have shown here, a search for fuzzy complexes involved in one-to-many and many-to-one signaling could shed new light on these novel and interesting protein-protein "flexible nets".

Molecular recognition features (MoRFs) are intrinsically disordered protein regions that bind to partners via disorder-to-order transitions. In one-to-many binding, a single MoRF binds to two or more different partners individually. MoRF-based one-to-many protein-protein interaction examples were collected from the Protein Data Bank (PDB), yielding 23 MoRFs bound to 2 to 9 partners, with all pairs of same-MoRF partners having less than 25% sequence identity. Of these, 8 MoRFs were bound to 2 to 9 partners having completely different folds, while 15 MoRFs were bound to 2 to 5 partners having the same folds but with low sequence identities. For both types of partner variation, backbone and side chain torsion angle rotations were used to bring about the conformational changes needed to enable close fits between a single MoRF and distinct partners. Alternative splicing events (ASEs) and posttranslational modifications (PTMs) were also found to contribute to distinct partner binding. Since ASEs and PTMs both commonly occur in disordered regions, and since both ASEs and PTMs are often tissue-

specific, these data suggest that MoRFs, ASEs, and PTMs may collaborate to alter protein-protein interaction networks in different cell types. These data enlarge the set of carefully studied MoRFs that use inherent flexibility and that also use ASE-based and/or PTM-based surface modifications to enable the same disordered segment to selectively associate with two or more partners. The small number of residues involved in MoRFs and in their modifications by ASEs or PTMs may simplify the evolvability of signaling network diversity.

The binding sites on the structured partners may also bind additional disordered sequences that have amino acid substitutions (e.g. many-to-one signaling). If such complexes exist, it would be interesting to determine whether the amino acid changes in the MoRFs compensate for the already observed amino acid changes in the structured binding partners. We have a collection of many-to-one examples, so we can search this set of interactions to determine if any of the structured proteins in the many-to-one collection match any of the structured partners discussed herein. Such a finding would not only provide information about possible mutation compensation across protein-protein interaction interfaces involving disordered protein regions, but would also suggest new concepts with regard to the structural basis of protein-protein interaction networks.

The independent and overlapping binding profiles we observed in the many-to-one set gave us a novel way to look at the dynamics of structural binding sites with binding diversity like 14-3-3. Although it is a challenge for us to categorize those overlapping binding sites into subgroups (e.g. similar and intersecting) quantitatively, volume overlap ratio calculations seems to work best compared to methods such as

RMSD measure of MoRFs, sequence similarity of MoRFs and structure alignment score of the MoRF-domain complexes.

There are still many other interesting ideas waiting for us to try and test on this valuable many-to-one binding dataset. For example, examining the structures of MoRF pairs with different overlap ratios is what we will do in the next step. We are curious to see the pairwise secondary structure profiles within different overlap ratio groups (e.g. similar and intersecting). Partner selection of MoRF characterized by diverse binding is another interesting subject to follow up. Although we have observed posttranslation modifications (PTMs) and alternatively spliced events (ASE) with regard to the MoRFs in our one-to-many set and proposed that the three major players (MoRFs, PTMs and ASEs) contribute significantly to the highly complex protein interaction networks in eukaryote cells, establishing a more systematically computational approach is necessary to validate our hypothesis.

Our results contribute to a better understanding of the role of disorder binding regions (MoRFs) that may serve as protein interaction hubs. Exploring the diverse binding partners of our collected MoRF sets and the corresponding complex conformations definitely give us a general Rosetta stone to interpret the underlying biological mechanisms and evolutional aptness. The importance and indispensability of hub proteins is apparent as they appear to evolve more slowly and are more likely to be vital for survival. Given their importance, many human disease-associated proteins related to cancer, diabetes, autoimmune disease, neurodegenerative disease and cardiovascular disease are found to have predicted disordered binding regions (MoRFs) as we expect [121]. These MoRFs associate with other structured partners and considered

as promising druggable interactions because of their high specificity and low affinity for binding. Since intrinsic protein disorder have high tendency to participate directly in large numbers of pairwise protein-protein interactions, these promiscuous protein interactions usually are toxic when overexpressed. They are dosage sensitivity. The fact is contrast to knockout or knockdown model, indicating there is something special about the excess participation specifically in pairwise interactions [122].

Binding with relatively low affinity is an advantageous attribute for transient, conditional and tunable interactions which is needed for many regulatory events. Therefore, this study will help pave the way for the development of novel pathways by designing intervening disordered peptides or small molecule having binding potential for particular partners but with tighter binding affinity.

REFERENCES

1. Sickmeier M, Hamilton JA, LeGall T, Vacic V, Cortese MS, Tantos A, Szabo B, Tompa P, Chen J, Uversky VN, Obradovic Z, Dunker AK. (2007) DisProt: the Database of Disordered Proteins. *Nucleic acids research* **35**, D786-793.

2. Xue B, Dunker AK, Uversky VN. (2012) Orderly order in protein intrinsic disorder distribution: disorder in 3500 proteomes from viruses and the three domains of life. *Journal of biomolecular structure & dynamics* **30**, 137-149.

3. Dunker AK, Brown CJ, Lawson JD, Iakoucheva LM, Obradovic Z. (2002) Intrinsic disorder and protein function. *Biochemistry* **41**, 6573-6582.

4. Dunker AK, Brown CJ, Obradovic Z. (2002) Identification and functions of usefully disordered proteins. *Adv Protein Chem* **62**, 25-49.

5. Vucetic S, Xie H, Iakoucheva LM, Oldfield CJ, Dunker AK, Obradovic Z, Uversky VN. (2007) Functional anthology of intrinsic disorder. 2. Cellular components, domains, technical terms, developmental processes, and coding sequence diversities correlated with long disordered regions. *J Proteome Res* **6**, 1899-1916.

6. Xie H, Vucetic S, Iakoucheva LM, Oldfield CJ, Dunker AK, Obradovic Z, Uversky VN. (2007) Functional anthology of intrinsic disorder. 3. Ligands, post-translational modifications, and diseases associated with intrinsically disordered proteins. *J Proteome Res* **6**, 1917-1932.

7. Xie H, Vucetic S, Iakoucheva LM, Oldfield CJ, Dunker AK, Uversky VN, Obradovic Z. (2007) Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions. *J Proteome Res* **6**, 1882-1898.

8. Fukuchi S, Hosoda K, Homma K, Gojobori T, Nishikawa K. (2011) Binary classification of protein molecules into intrinsically disordered and ordered segments. *Bmc Struct Biol* **11**.

9. Romero PR, Zaidi S, Fang YY, Uversky VN, Radivojac P, Oldfield CJ, Cortese MS, Sickmeier M, LeGall T, Obradovic Z, Dunker AK. (2006) Alternative splicing in concert with protein intrinsic disorder enables increased functional diversity in multicellular organisms. *Proc Natl Acad Sci U S A* **103**, 8390-8395.

10. Jeong H, Mason SP, Barabasi AL, Oltvai ZN. (2001) Lethality and centrality in protein networks. *Nature* **411**, 41-42.

11. Barabasi AL, Oltvai ZN. (2004) Network biology: understanding the cell's functional organization. *Nat Rev Genet* **5**, 101-113.

12. Hasty J, Collins JJ. (2001) Protein interactions. Unspinning the web. *Nature* **411**, 30-31.

13. Pauling L. (1940) A Theory of the Structure and Process of Formation of Antibodies*. *Journal of the American Chemical Society* **62**, 2643-2657.

14. Dunker AK, Garner E, Guilliot S, Romero P, Albrecht K, Hart J, Obradovic Z, Kissinger C, Villafranca JE. (1998) Protein disorder and the evolution of molecular recognition: theory, predictions and observations. *Pac Symp Biocomput*, 473-484.

15. Kriwacki RW, Hengst L, Tennant L, Reed SI, Wright PE. (1996) Structural studies of p21Waf1/Cip1/Sdi1 in the free and Cdk2-bound state: conformational disorder mediates binding diversity. *Proceedings of the National Academy of Sciences of the United States of America* **93**, 11504-11509.

16. James LC, Roversi P, Tawfik DS. (2003) Antibody multispecificity mediated by conformational diversity. *Science* **299**, 1362-1367.

17. Patil A, Kinoshita K, Nakamura H. (2010) Hub promiscuity in protein-protein interaction networks. *Int J Mol Sci* **11**, 1930-1943.

18. Kim PM, Sboner A, Xia Y, Gerstein M. (2008) The role of disorder in interaction networks: a structural analysis. *Mol Syst Biol* **4**, 179.

19. Haynes C, Oldfield CJ, Ji F, Klitgord N, Cusick ME, Radivojac P, Uversky VN, Vidal M, Iakoucheva LM. (2006) Intrinsic disorder is a common feature of hub proteins from four eukaryotic interactomes. *PLoS Comput Biol* **2**, e100.

20. Ekman D, Light S, Bjorklund AK, Elofsson A. (2006) What properties characterize the hub proteins of the protein-protein interaction network of Saccharomyces cerevisiae? *Genome Biol* **7**.

21. Patil A, Nakamura H. (2006) Disordered domains and high surface charge confer hubs with the ability to interact with multiple proteins in interaction networks. *Febs Lett* **580**, 2041-2045.

22. Dosztanyi Z, Chen J, Dunker AK, Simon I, Tompa P. (2006) Disorder and sequence repeats in hub proteins and their implications for network evolution. *J Proteome Res* **5**, 2985-2995.

23. Boxem M, Maliga Z, Klitgord N, Li N, Lemmens I, Mana M, de Lichtervelde L, Mul JD, van de Peut D, Devos M, Simonis N, Yildirim MA, Cokol M, Kao HL, de Smet AS, Wang HD, Schlaitz AL, Hao T, Milstein S, Fan CY, Tipsword M, Drew K, Galli M, Rhrissorrakrai K, Drechsel D, Koller D, Roth FP, Iakoucheva LM, Dunker AK, Bonneau R, Gunsalus KC, Hill DE, Piano F, Tavernier J, van den Heuvel S, Hyman AA, Vidal M. (2008) A protein domain-based interactome network for C-elegans early embryogenesis. *Cell* **134**, 534-545.

24. Singh GP, Dash D. (2007) Intrinsic disorder in yeast transcriptional regulatory network. *Proteins* **68**, 602-605.

25. Singh GP, Ganapathi M, Dash D. (2007) Role of intrinsic disorder in transient interactions of hub proteins. *Proteins* **66**, 761-765.

26. Kim PM, Sboner A, Xia Y, Gerstein M. (2008) The role of disorder in interaction networks: a structural analysis. *Molecular Systems Biology* **4**.

27. Bjorklund AK, Light S, Hedin L, Elofsson A. (2008) Quantitative assessment of the structural bias in protein-protein interaction assays. *Proteomics* **8**, 4657-4667.

28. Higurashi M, Ishida T, Kinoshita K. (2008) Identification of transient hub proteins and the possible structural basis for their multiple interactions. *Protein science : a publication of the Protein Society* **17**, 72-78.

29. Kahali B, Ahmad S, Ghosh TC. (2009) Exploring the evolutionary rate differences of party hub and date hub proteins in Saccharomyces cerevisiae protein-protein interaction network. *Gene* **429**, 18-22.

30. Manna B, Bhattacharya T, Kahali B, Ghosh TC. (2009) Evolutionary constraints on hub and non-hub proteins in human protein interaction network: insight from protein connectivity and intrinsic disorder. *Gene* **434**, 50-55.

31. Patil A, Kinoshita K, Nakamura H. (2010) Domain distribution and intrinsic disorder in hubs in the human protein-protein interaction network. *Protein science : a publication of the Protein Society* **19**, 1461-1468.

32. Puntervoll P, Linding R, Gemund C, Chabanis-Davidson S, Mattingsdal M, Cameron S, Martin DM, Ausiello G, Brannetti B, Costantini A, Ferre F, Maselli V, Via A, Cesareni G, Diella F, Superti-Furga G, Wyrwicz L, Ramu C, McGuigan C, Gudavalli R, Letunic I, Bork P, Rychlewski L, Kuster B, Helmer-Citterich M, Hunter WN, Aasland R, Gibson TJ. (2003) ELM server: A new resource for investigating short functional sites in modular eukaryotic proteins. *Nucleic acids research* **31**, 3625-3630.

33. Gould CM, Diella F, Via A, Puntervoll P, Gemund C, Chabanis-Davidson S, Michael S, Sayadi A, Bryne JC, Chica C, Seiler M, Davey NE, Haslam N, Weatheritt RJ, Budd A, Hughes T, Pas J, Rychlewski L, Trave G, Aasland R, Helmer-Citterich M, Linding R, Gibson TJ. (2010) ELM: the status of the 2010 eukaryotic linear motif resource. *Nucleic acids research* **38**, D167-180.

34. Fuxreiter M, Tompa P, Simon I. (2007) Local structural disorder imparts plasticity on linear motifs. *Bioinformatics* **23**, 950-956.

35. Davey NE, Shields DC, Edwards RJ. (2006) SLiMDisc: short, linear motif discovery, correcting for common evolutionary descent. *Nucleic acids research* **34**, 3546-3554.

36. Edwards RJ, Davey NE, Shields DC. (2007) SLiMFinder: a probabilistic method for identifying over-represented, convergently evolved, short linear motifs in proteins. *PLoS One* **2**, e967.

37. Callaghan AJ, Aurikko JP, Ilag LL, Gunter Grossmann J, Chandran V, Kuhnel K, Poljak L, Carpousis AJ, Robinson CV, Symmons MF, Luisi BF. (2004) Studies of the RNA degradosome-organizing domain of the Escherichia coli ribonuclease RNase E. *J Mol Biol* **340**, 965-979.

38. Mohan A, Oldfield CJ, Radivojac P, Vacic V, Cortese MS, Dunker AK, Uversky VN. (2006) Analysis of molecular recognition features (MoRFs). *J Mol Biol* **362**, 1043-1059.

39. Obenauer JC, Yaffe MB. (2004) Computational prediction of protein-protein interactions. *Methods Mol Biol* **261**, 445-468.

40. Valencia A, Pazos F (2008) Computational Methods to Predict Protein Interaction Partners in *Protein-protein Interactions and Networks* (Panchenko, A. & Przytycka, T. M., eds) pp. 67-81, Springer.

41. Obenauer JC, Cantley LC, Yaffe MB. (2003) Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic acids research* **31**, 3635-3641.

42. Kadaveru K, Vyas J, Schiller MR. (2008) Viral infection and human disease--insights from minimotifs. *Frontiers in bioscience : a journal and virtual library* **13**, 6455-6471.

43. Mi T, Merlin JC, Deverasetty S, Gryk MR, Bill TJ, Brooks AW, Lee LY, Rathnayake V, Ross CA, Sargeant DP, Strong CL, Watts P, Rajasekaran S, Schiller MR. (2012) Minimotif Miner 3.0: database expansion and significantly improved reduction of false-positive predictions from consensus sequences. *Nucleic acids research* **40**, D252-260.

44. Garner E, Romero P, Dunker AK, Brown C, Obradovic Z. (1999) Predicting Binding Regions within Disordered Proteins. *Genome Inform Ser Workshop Genome Inform* **10**, 41-50.

45. Oldfield CJ, Cheng Y, Cortese MS, Romero P, Uversky VN, Dunker AK. (2005) Coupled folding and binding with alpha-helix-forming molecular recognition elements. *Biochemistry* **44**, 12454-12470.

46. Oldfield CJ, Meng J, Yang JY, Yang MQ, Uversky VN, Dunker AK. (2008) Flexible nets: disorder and induced fit in the associations of p53 and 14-3-3 with their partners. *BMC Genomics* **9 Suppl 1**, S1.

47. Dosztanyi Z, Meszaros B, Simon I. (2009) ANCHOR: web server for predicting protein binding regions in disordered proteins. *Bioinformatics* **25**, 2745-2746.

48. Disfani FM, Hsu WL, Mizianty MJ, Oldfield CJ, Xue B, Dunker AK, Uversky VN, Kurgan L. (2012) MoRFpred, a computational tool for sequence-based prediction and characterization of short disorder-to-order transitioning binding regions in proteins. *Bioinformatics* **28**, i75-83.

49. Bourhis J-M, Johansson K, Receveur-Brechot V, Oldfield CJ, Dunker KA, Canard B, Longhi S. (2004) The C-terminal domain of measles virus nucleoprotein belongs to the class of intrinsically disordered proteins that fold upon binding to their physiological partner. *Virus Res* **99**, 157-167.

50. Gunasekaran K, Tsai CJ, Nussinov R. (2004) Analysis of ordered and disordered protein complexes reveals structural features discriminating between stable and unstable monomers. *J Mol Biol* **341**, 1327-1341.

51. Dunker AK, Cortese MS, Romero P, Iakoucheva LM, Uversky VN. (2005) Flexible nets. The roles of intrinsic disorder in protein interaction networks. *FEBS J* **272**, 5129-5148.

52. Patil A, Kinoshita K, Nakamura H. (2010) Hub promiscuity in protein-protein interaction networks. *International journal of molecular sciences* **11**, 1930-1943.

53. Lowe ED, Tews I, Cheng KY, Brown NR, Gul S, Noble ME, Gamblin SJ, Johnson LN. (2002) Specificity determinants of recruitment peptides bound to phospho-CDK2/cyclin A. *Biochemistry* **41**, 15625-15634.

54. Avalos JL, Celic I, Muhammad S, Cosgrove MS, Boeke JD, Wolberger C. (2002) Structure of a Sir2 enzyme bound to an acetylated p53 peptide. *Molecular cell* **10**, 523-535.

55. Mujtaba S, He Y, Zeng L, Yan S, Plotnikova O, Sachchidanand, Sanchez R, Zeleznik-Le NJ, Ronai Z, Zhou MM. (2004) Structural mechanism of the bromodomain of the coactivator CBP in p53 transcriptional activation. *Molecular cell* **13**, 251-263.

56. Rustandi RR, Baldisseri DM, Weber DJ. (2000) Structure of the negative regulatory domain of p53 bound to S100B(betabeta). *Nature structural biology* **7**, 570-574.

57. van Heusden GP. (2005) 14-3-3 proteins: regulators of numerous eukaryotic proteins. *IUBMB Life* **57**, 623-629.

58. Bustos DM, Iglesias AA. (2006) Intrinsic disorder is a key characteristic in partners that bind 14-3-3 proteins. *Proteins-Structure Function and Bioinformatics* **63**, 35-42.

59. Bustos DM. (2012) The role of protein disorder in the 14-3-3 interaction network. *Mol Biosyst* **8**, 178-184.

60. Mackintosh C. (2004) Dynamic interactions between 14-3-3 proteins and phosphoproteins regulate diverse cellular processes. *Biochem J* **381**, 329-342.

61. Neduva V, Russell RB. (2006) Peptides mediating interaction networks: new leads at last. *Curr Opin Biotech* **17**, 465-471.

62. London N, Movshovitz-Attias D, Schueler-Furman O. (2010) The structural basis of peptide-protein binding strategies. *Structure* **18**, 188-199.

63. Zucconi A, Panni S, Paoluzi S, Castagnoli L, Dente L, Cesareni G. (2000) Domain repertoires as a tool to derive protein recognition rules. *Febs Lett* **480**, 49-54.

64. Dyson HJ, Wright PE. (2005) Intrinsically unstructured proteins and their functions. *Nature reviews Molecular cell biology* **6**, 197-208.

65. Rittinger K, Budman J, Xu J, Volinia S, Cantley LC, Smerdon SJ, Gamblin SJ, Yaffe MB. (1999) Structural analysis of 14-3-3 phosphopeptide complexes identifies a dual role for the nuclear export signal of 14-3-3 in ligand binding. *Molecular cell* **4**, 153-166.

66. Uversky VN, Dunker AK. (2010) Understanding protein non-folding. *Biochim Biophys Acta* **1804**, 1231-1264.

67. Wright PE, Dyson HJ. (1999) Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J Mol Biol* **293**, 321-331.

68. Dunker AK, Lawson JD, Brown CJ, Williams RM, Romero P, Oh JS, Oldfield CJ, Campen AM, Ratliff CM, Hipps KW, Ausio J, Nissen MS, Reeves R, Kang C, Kissinger CR, Bailey RW, Griswold MD, Chiu W, Garner EC, Obradovic Z. (2001) Intrinsically disordered protein. *J Mol Graph Model* **19**, 26-59.

69. Tompa P. (2002) Intrinsically unstructured proteins. *Trends in biochemical sciences* **27**, 527-533.

70. Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT. (2004) Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J Mol Biol* **337**, 635-645.

71. Hsu WL, Oldfield C, Meng J, Huang F, Xue B, Uversky VN, Romero P, Dunker AK. (2012) Intrinsic protein disorder and protein-protein interactions. *Pac Symp Biocomput*, 116-127.

72. Kriwacki RW, Hengst L, Tennant L, Reed SI, Wright PE. (1996) Structural studies of p21Waf1/Cip1/Sdi1 in the free and Cdk2-bound state: conformational disorder mediates binding diversity. *Proc Natl Acad Sci U S A* **93**, 11504-11509.

73. Carugo O, Argos P. (1997) Protein-protein crystal-packing contacts. *Protein science : a publication of the Protein Society* **6**, 2261-2263.

74. Henrick K, Thornton JM. (1998) PQS: a protein quaternary structure file server. *Trends in biochemical sciences* **23**, 358-361.

75. Shindyalov IN, Bourne PE. (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng* **11**, 739-747.

76. Shatsky M, Nussinov R, Wolfson HJ. (2004) A method for simultaneous alignment of multiple protein structures. *Proteins* **56**, 143-156.

77. Simossis VA, Heringa J. (2005) PRALINE: a multiple sequence alignment toolbox that integrates homology-extended and secondary structure information. *Nucleic acids research* **33**, W289-294.

78. Oldfield CJ, Meng J, Yang JY, Uversky VN, Dunker AK. (2007) Intrinsic Disorder in Protein-Protein Interaction Networks: Case Studies of Complexes Involving p53 and 14-3-3. *BIOCOMP'07*, 553-566.

79. Doolittle RF (1986) *Of Urfs and Orfs: A Primer on How to Analyze Derived Amino Acid Sequences*, University Science Books, Mill Valley, California.

80. Grishin NV, Phillips MA. (1994) The subunit interfaces of oligomeric enzymes are conserved to a similar extent to the overall protein sequences. *Protein Sci* **3**, 2455-2458.

81. Wang L, Zuercher WJ, Consler TG, Lambert MH, Miller AB, Orband-Miller LA, McKee DD, Willson TM, Nolte RT. (2006) X-ray crystal structures of the estrogen-related receptor-gamma ligand binding domain in three functional states reveal the molecular basis of small molecule regulation. *J Biol Chem* **281**, 37773-37781.

82. Madauss KP, Grygielko ET, Deng SJ, Sulpizio AC, Stanley TB, Wu C, Short SA, Thompson SK, Stewart EL, Laping NJ, Williams SP, Bray JD. (2007) A structural and in vitro characterization of asoprisnil: A selective progesterone receptor modulator. *Mol Endocrinol* **21**, 1066-1081.

83. Xu HE, Stanley TB, Montana VG, Lambert MH, Shearer BG, Cobb JE, McKee DD, Galardi CM, Plunket KD, Nolte RT, Parks DJ, Moore JT, Kliewer SA, Willson TM, Stimmel JB. (2002) Structural basis for antagonist-mediated recruitment of nuclear co-repressors by PPAR alpha. *Nature* **415**, 813-817.

84. Darimont BD, Wagner RL, Apriletti JW, Stallcup MR, Kushner PJ, Baxter JD, Fletterick RJ, Yamamoto KR. (1998) Structure and specificity of nuclear receptor-coactivator interactions. *Genes & development* **12**, 3343-3356.

85. Shiau AK, Barstad D, Radek JT, Meyers MJ, Nettles KW, Katzenellenbogen BS, Katzenellenbogen JA, Agard DA, Greene GL. (2002) Structural characterization of a subtype-selective ligand reveals a novel mode of estrogen receptor antagonism. *Nature structural biology* **9**, 359-364.

86. Xu RX, Lambert MH, Wisely BB, Warren EN, Weinert EE, Waitt GM, Williams JD, Collins JL, Moore LB, Willson TM, Moore JT. (2004) A structural basis for constitutive activity in the human CAR/RXRalpha heterodimer. *Molecular cell* **16**, 919-928.

87. Estebanez-Perpina E, Moore JM, Mar E, Delgado-Rodrigues E, Nguyen P, Baxter JD, Buehrer BM, Webb P, Fletterick RJ, Guy RK. (2005) The molecular mechanisms of coactivator utilization in ligand-dependent transactivation by the androgen receptor. *J Biol Chem* **280**, 8060-8068.

88. Soisson SM, Parthasarathy G, Adams AD, Sahoo S, Sitlani A, Sparrow C, Cui J, Becker JW. (2008) Identification of a potent synthetic FXR agonist with an unexpected mode of binding and activation. *Proc Natl Acad Sci U S A* **105**, 5337-5342.

89. Xue Y, Chao E, Zuercher WJ, Willson TM, Collins JL, Redinbo MR. (2007) Crystal structure of the PXR-T1317 complex provides a scaffold to examine the potential for receptor antagonism. *Bioorg Med Chem* **15**, 2156-2166.

90. Rosonina E, Blencowe BJ. (2004) Analysis of the requirement for RNA polymerase II CTD heptapeptide repeats in pre-mRNA splicing and 3'-end cleavage. *RNA* **10**, 581-589.

91. Zhang Y, Kim Y, Genoud N, Gao J, Kelly JW, Pfaff SL, Gill GN, Dixon JE, Noel JP. (2006) Determinants for dephosphorylation of the RNA polymerase II C-terminal domain by Scp1. *Molecular cell* **24**, 759-770.

92. Meinhart A, Cramer P. (2004) Recognition of RNA polymerase II carboxy-terminal domain by 3'-RNA-processing factors. *Nature* **430**, 223-226.

93. Fabrega C, Shen V, Shuman S, Lima CD. (2003) Structure of an mRNA capping enzyme bound to the phosphorylated carboxy-terminal domain of RNA polymerase II. *Molecular cell* **11**, 1549-1561.

94. Huang Y, Fang J, Bedford MT, Zhang Y, Xu RM. (2006) Recognition of histone H3 lysine-4 methylation by the double tudor domain of JMJD2A. *Science* **312**, 748-751.

95. Ooi SK, Qiu C, Bernstein E, Li K, Jia D, Yang Z, Erdjument-Bromage H, Tempst P, Lin SP, Allis CD, Cheng X, Bestor TH. (2007) DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA. *Nature* **448**, 714-717.

96. Ruthenburg AJ, Wang W, Graybosch DM, Li H, Allis CD, Patel DJ, Verdine GL. (2006) Histone H3 recognition and presentation by the WDR5 module of the MLL1 complex. *Nat Struct Mol Biol* **13**, 704-712.

97. Ramon-Maiques S, Kuo AJ, Carney D, Matthews AG, Oettinger MA, Gozani O, Yang W. (2007) The plant homeodomain finger of RAG2 recognizes histone H3 methylated at both lysine-4 and arginine-2. *Proc Natl Acad Sci U S A* **104**, 18993-18998.

98. Forneris F, Binda C, Adamo A, Battaglioli E, Mattevi A. (2007) Structural basis of LSD1-CoREST selectivity in histone H3 recognition. *J Biol Chem* **282**, 20070-20074.

99. Clements A, Poux AN, Lo WS, Pillus L, Berger SL, Marmorstein R. (2003) Structural basis for histone and phosphohistone binding by the GCN5 histone acetyltransferase. *Molecular cell* **12**, 461-473.

100. Macdonald N, Welburn JP, Noble ME, Nguyen A, Yaffe MB, Clynes D, Moggs JG, Orphanides G, Thomson S, Edmunds JW, Clayton AL, Endicott JA, Mahadevan LC. (2005) Molecular basis for the recognition of phosphorylated and phosphoacetylated histone h3 by 14-3-3. *Molecular cell* **20**, 199-211.

101. Couture JF, Collazo E, Ortiz-Tello PA, Brunzelle JS, Trievel RC. (2007) Specificity and mechanism of JMJD2A, a trimethyllysine-specific histone demethylase. *Nat Struct Mol Biol* **14**, 689-695.

102. Zhang X, Yang Z, Khan SI, Horton JR, Tamaru H, Selker EU, Cheng X. (2003) Structural basis for the product specificity of histone lysine methyltransferases. *Molecular cell* **12**, 177-185.

103. Cheng Y, Oldfield CJ, Meng J, Romero P, Uversky VN, Dunker AK. (2007) Mining alpha-helix-forming molecular recognition features with cross species sequence alignments. *Biochemistry* **46**, 13468-13477.

104. Fares MA, Ruiz-Gonzalez MX, Labrador JP. (2011) Protein coadaptation and the design of novel approaches to identify protein-protein interactions. *IUBMB Life* **63**, 264-271.

105. Diella F, Haslam N, Chica C, Budd A, Michael S, Brown NP, Trave G, Gibson TJ. (2008) Understanding eukaryotic linear motifs and their role in cell signaling and regulation. *Frontiers in Bioscience* **13**, 6580-6603.

106. Gfeller D, Butty F, Wierzbicka M, Verschueren E, Vanhee P, Huang HM, Ernst A, Dar N, Stagljar I, Serrano L, Sidhu SS, Bader GD, Kim PM. (2011) The multiple-specificity landscape of modular peptide recognition domains. *Molecular Systems Biology* **7**.

107. Davey NE, Van Roey K, Weatheritt RJ, Toedt G, Uyar B, Altenberg B, Budd A, Diella F, Dinkel H, Gibson TJ. (2012) Attributes of short linear motifs. *Mol Biosyst* **8**, 268-281.

108. Dinkel H, Michael S, Weatheritt RJ, Davey NE, Van Roey K, Altenberg B, Toedt G, Uyar B, Seiler M, Budd A, Jodicke L, Dammert MA, Schroeter C, Hammer M, Schmidt T, Jehl P, McGuigan C, Dymecka M, Chica C, Luck K, Via A, Chatr-Aryamontri A, Haslam N, Grebnev G, Edwards RJ, Steinmetz MO, Meiselbach H, Diella F, Gibson TJ. (2012) ELM--the database of eukaryotic linear motifs. *Nucleic acids research* **40**, D242-251.

109. Buljan M, Chalancon G, Eustermann S, Wagner GP, Fuxreiter M, Bateman A, Babu MM. (2012) Tissue-specific splicing of disordered segments that embed binding motifs rewires protein interaction networks. *Molecular cell* **46**, 871-883.

110. Ellis JD, Barrios-Rodiles M, Colak R, Irimia M, Kim T, Calarco JA, Wang X, Pan Q, O'Hanlon D, Kim PM, Wrana JL, Blencowe BJ. (2012) Tissue-specific alternative splicing remodels protein-protein interaction networks. *Molecular cell* **46**, 884-892.

111. Iakoucheva LM, Radivojac P, Brown CJ, O'Connor TR, Sikes JG, Obradovic Z, Dunker AK. (2004) The importance of intrinsic disorder for protein phosphorylation. *Nucleic acids research* **32**, 1037-1049.

112. Gao J, Thelen JJ, Dunker AK, Xu D. (2010) Musite, a tool for global prediction of general and kinase-specific phosphorylation sites. *Mol Cell Proteomics* **9**, 2586-2600.

113. Gao J, Xu D. (2012) Correlation between posttranslational modification and intrinsic disorder in protein. *Pac Symp Biocomput*, 94-103.

114. Vacic V, Uversky VN, Dunker AK, Lonardi S. (2007) Composition Profiler: a tool for discovery and visualization of amino acid composition differences. *BMC bioinformatics* **8**, 211.

115. Xue B, Dunker AK, Uversky VN. (2010) Retro-MoRFs: Identifying Protein Binding Sites by Normal and Reverse Alignment and Intrinsic Disorder Prediction. *Int J Mol Sci* **11**, 3725-3747.

116. Mahoney NM, Rozwarski DA, Fedorov E, Fedorov AA, Almo SC. (1999) Profilin binds proline-rich ligands in two distinct amide backbone orientations. *Nature structural biology* **6**, 666-671.

117. Hsu WL, Oldfield CJ, Xue B, Meng J, Romero P, Uversky VN, Dunker AK. (2013) Exploring the binding diversity of intrinsically disordered proteins involved in one-to-many signaling. *Protein Science* **22**, 258-273.

118. Spolar RS, Record MT, Jr. (1994) Coupling of local folding to site-specific binding of proteins to DNA. *Science* **263**, 777-784.

119. Espinoza-Fonseca LM. (2011) Aromatic residues link binding and function of intrinsically disordered proteins. *Molecular BioSystems*.

120. Tompa P, Fuxreiter M. (2008) Fuzzy complexes: polymorphism and structural disorder in protein-protein interactions. *Trends Biochem Sci* **33**, 2-8.

121. Cheng Y, LeGall T, Oldfield CJ, Mueller JP, Van YY, Romero P, Cortese MS, Uversky VN, Dunker AK. (2006) Rational drug design via intrinsically disordered protein. *Trends Biotechnol* **24**, 435-442.

122. Marcotte EM, Tsechansky M. (2009) Disorder, Promiscuity, and Toxic Partnerships. *Cell* **138**, 16-18.

CURRICULUM VITAE

Wei-Lun Hsu

Education

Doctor of Philosophy, Biochemistry and Molecular Biology
Minor: Life Science
July 2013
Indiana University, Indianapolis, IN, USA

Master of Science, Bioinformatics and Systems Biology
June 2005
National Chiao Tung University, Hsinchu, Taiwan

Bachelor of Science, Public Health
Bachelor of Science, Information Management
June 2003
National Taiwan University, Taipei, Taiwan

Scientific Discipline and Research Interests

Bioinformatics, computational biology, systems biology, data mining, machine learning, big data analysis, biostatistics, networks biology, protein structure prediction and modeling, structural biology, genomics/proteomic data analysis and visualization, sequence/structure/function analysis, biophysics, molecular dynamics simulations, signaling pathway, biomedical database, disease informatics, drug target identification and drug discovery.

Professional Experience and Association Memberships

Member of Advancing Science, Serving Society, 2011-2013.
Member of Biophysical Society, 2009-2011.
Reviewer for "PROTEINS: Structure, Function, and Bioinformatics", 2008.

Honors, Awards, Fellowships

Double Major in National Taiwan University (1998~2003)
Math and Science Talented Program in Taipei First Girls High School (1995~1998)
Spirit Award in Biology in Taipei Science Fair (1993)

Specialties and Skills

Programming Languages: Perl, C/C++, CGI, R, SAS, MATLAB, Mathematica, PHP, MySQL, HTML, JavaScript, CSS, XML, Python, C#, ASP.NET, UML, Java.

Statistics Skills: Data analysis, interpretation and visualization, statistical inference, sampling, hypothesis testing, probability theory, ANOVA variance analysis, correlation analysis, regression model, cluster analysis, machine learning and data mining.

Bioinformatics Techniques: Sequence alignment, genome annotation, computational evolutionary biology, phylogenetic analysis, regular expression, pattern recognition, pathway analysis, motif and profile analysis, protein structure comparison, alignment, prediction and classification.

Bioinformatics Databases and Tools: NCBI, EMBL, Ensembl, UCSC Genome Browser, ExPASy, dbSNP, OMIM, UniProt, PIR, Swiss-Prot, Pfam, PROSITE, DIP, KEGG, PDB, SCOP, CATH, GO, BLAST, PyMOL, RasMol, Swiss PDB Viewer, Ingenuity Pathway Analysis, LIBSVM, Weka, molecular dynamics simulation, AMBER, SYBYL, VMD.

Operating Systems: Microsoft Windows, Unix/Linux, Mac OS X.

Research and Work Experience

Center for Computational Biology and Bioinformatics, Indiana University
Research Assistant under Dr. A. Keith Dunker (Aug 2008~July 2013)

- Established molecular recognition feature (MoRF) databases and investigated the underlying regulatory mechanisms including alternative splicing events and posttranslational modifications.
- Explored the binding diversity of intrinsically disordered proteins and integrated various bioinformatics algorithms, sequence/structure analysis and visualization tools to study the motif patterns and conformational changes of binding interfaces within protein interaction networks.
  (our publication was selected as the editorial focus and video highlight in protein science, http://youtu.be/eoGumjI9zBw)
- Developed novel approaches to identify biomarkers or druggable targets in order to intervene disease signaling pathways by rational "*unstructure-based drug design*" and large scale genomics/proteomics data analyses.

Center for Computational Biology and Bioinformatics, Indiana University
Rotation student under Dr. Samy Meroueh (Jan 2008~Mar 2008)

- Carried out molecular dynamics simulations and Molecular Mechanic/Poisson-Boltzmann Surface Area (MM-PBSA) analysis to investigate protein-protein interactions. Participated in protein interaction interface re-design.

Center for Computational Biology and Bioinformatics, Indiana University
Rotation student under Dr. Yaoqi Zhou (Oct 2007~Dec 2007)

- Developed a robotic algorithm to build a representative structural domain database based on sequence comparisons and domain assignment methods.

Institute of Bioinformatics and Systems Biology, National Chiao Tung University
Research Assistant in Molecular Bioinformatics Center (July 2005~Jun 2007)

- Developed/maintained/IT-supported online bioinformatics databases, applications and webservers in the molecular bioinformatics center and helped manage/propose research grants.

Institute of Bioinformatics and Systems Biology, National Chiao Tung University
Research Assistant under Dr. Jenn-Kang Hwang (Aug 2003~Jun 2005)

- Constructed statistical/mathematical models to predict protein relative solvent accessibility from amino acid sequence (by machine learning approaches e.g. support vector machine).

Adaptive Intelligent Internet Agents Lab, Institute of Information Science, Academia Sinica
Student under Dr. Chun-Nan Hsu (Jan 2003~Jul 2003)

- Applied the web wrapper agent toolbox to bioinformatics research.

Project Works
UML modeling language programmer (Jul 2002~Dec 2002)

- Annotated a knowledge information sharing system based on object-oriented software engineering.

Department of Health, Executive Yuan
Internship in Food and Drug Administration (Jul 2001~ Aug 2001)

- Case studied on drug abuse incidents and related healthcare/clinical data.

Division of Information Technology
Internship in China Times (Jun 2001~ Jul 2001)

- Interviewed, reported and edited information technology news as a journalist.

Conferences Abstracts, Posters and Presentations

1. Hsu WL, Dunker AK, et al. Mechanisms of binding diversity and partner selection in protein disorder: Molecular recognition features mediating protein interaction networks. **Accepted for oral presentations** at 2013 Boston Taiwanese Biotechnology Symposium, Cambridge, MA, June 15, 2013.
2. Kallenbach J, Hsu WL, Dunker AK, Alterovitz G. Order-Disorder Interface Characterization Reveals Critical Factors for Disease and Drug Targets**. Accepted for papers/podium presentations** at the American Medical Informatics Association (AMIA), San Francisco, CA, Mar 18-20, 2013.
3. Disfani FM, Hsu WL, Mizianty MJ, Oldfield CJ, Xue B, Dunker A, Uversky V, Kurgan L. MoRFpred, a computational tool for sequence-based prediction and characterization of disorder-to-order transitioning binding sites in proteins. Accepted for publication and proceedings talks at the meeting in July 15-17, 2012 in Intelligent Systems for Molecular Biology (ISMB) in Long Beach, California.
4. Hsu WL, Oldfield C, Meng J, Huang F, Xue B, Uversky VN, Dunker AK. Intrinsic Protein Disorder and Protein-Protein Interactions. Accepted for publication in the proceedings and oral presentation at the meeting in January 3-7, 2012 in Pacific Symposium on Biocomputing (PSB) in Big Island, Hawaii.

5. Hsu WL, Dunker AK, et al. Molecular recognition features facilitate the binding diversity of hub proteins. Biophysical Society (BPS) 55th Annual Meeting, Baltimore, MD, March 2011.
6. Hsu WL, Dunker AK, et al. Molecular recognition features from intrinsically disordered regions are druggable targets. Pacific Symposium on Biocomputing, Big Island, HI, January 2011.
7. Hsu WL, Dunker AK, et al. Exploring the binding diversity of intrinsic disorder. Biophysical Society (BPS) 54th Annual Meeting, San Francisco, CA, February 2010.

Publications

1. Hsu WL, Oldfield CJ, Eshel Faraggi, Xue B, Meng J, Huang F, Romero P, Uversky V, Dunker AK. Characterizing binding profiles of many-to-one disordered protein complexes. (In Progress).
2. Hsu WL, Oldfield CJ, Xue B, Meng J, Huang F, Romero P, Uversky V, Dunker AK. Exploring the binding diversity of intrinsically disordered proteins involved in one-to-many binding. Protein Science 22, 258-273 (2013).
3. Disfani FM, Hsu WL, Mizianty MJ, Oldfield CJ, Xue B, Dunker AK, Uversky V, Kurgan L. MoRFpred, a computational tool for sequence-based prediction and characterization of disorder-to-order transitioning binding sites in proteins. Bioinformatics. 2012; 28(12): i75-i83.
4. Hsu WL, Oldfield C, Meng J, Huang F, Xue B, Uversky VN, Romero P, Dunker AK. Intrinsic Protein Disorder and Protein-Protein Interactions. PSB 2012;17:116-27.
5. Huang F, Oldfield C, Meng J, Hsu WL, Xue B, Uversky VN, Romero P, Dunker AK. Subclassifying Disordered Proteins by the CH-CDF Plot Method. PSB 2012;17:128-39.
6. Xue B, Hsu WL, Lee JH, Lu H, Dunker AK, Uversky VN. SPA: Short peptide analyzer of intrinsic disorder status of short peptides. Genes Cells. 2010;15(6):635-46.
7. Liang S, Li L, Hsu WL, Pilcher MN, Uversky V, Zhou Y, Dunker AK, Meroueh SO. Exploring the molecular design of protein interaction sites with molecular dynamics simulations and free energy calculations. Biochemistry. 2009;48(2):399-414. PMCID: 2754190.
8. Lu CH, Huang SW, Lai YL, Lin CP, Shih CH, Huang CC, Hsu WL, Hwang JK. On the relationship between the protein structure and protein dynamics. Proteins-Structure Function and Bioinformatics. 2008;72(2):625-34.
9. Lin YS, Hsu WL, Hwang JK, Li WH. Proportion of solvent-exposed amino acids in a protein and rate of protein evolution. Mol Biol Evol. 2007;24(4):1005-11.